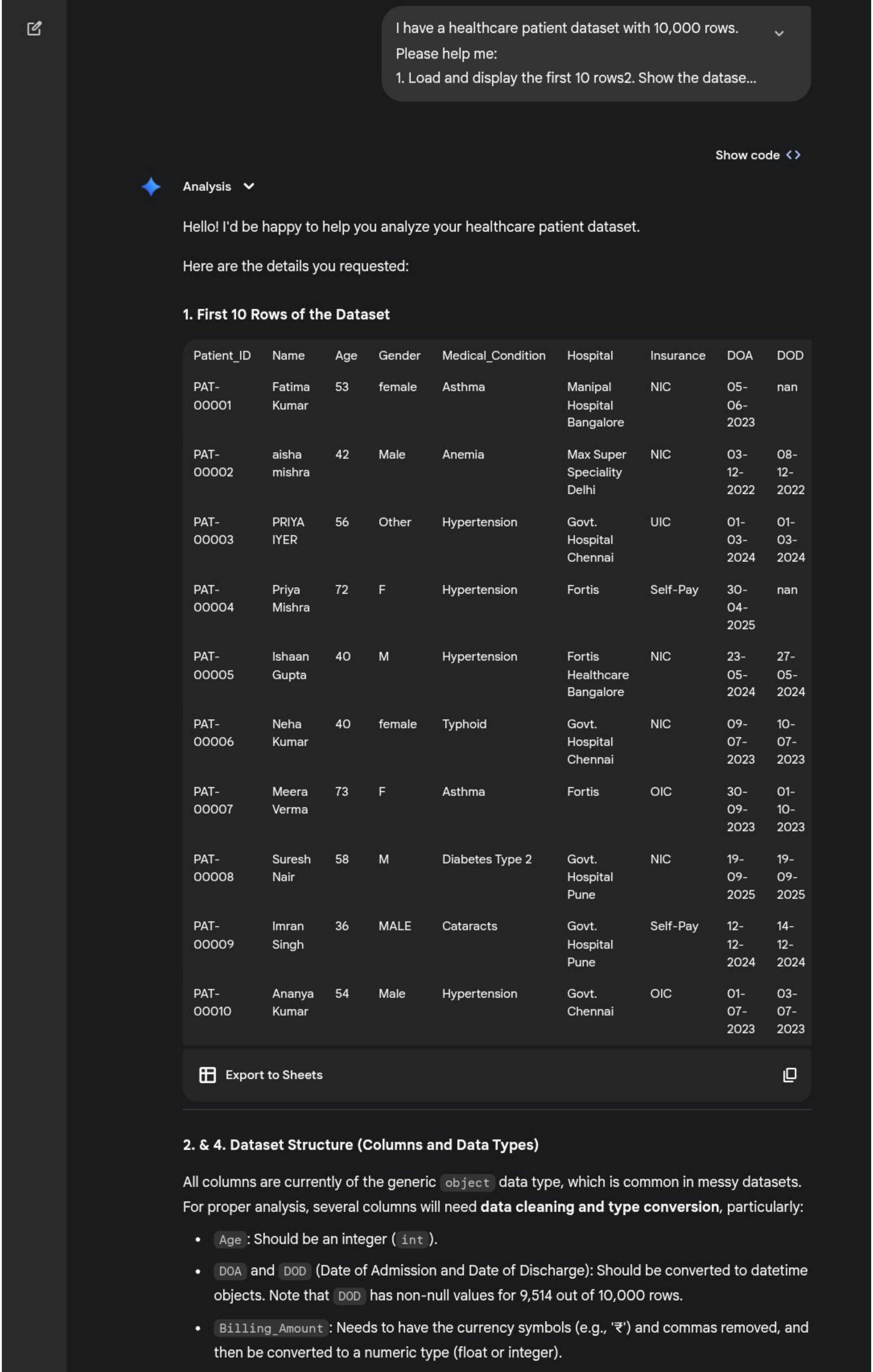


I have a healthcare patient dataset with 10,000 rows.
Please help me:
1. Load and display the first 10 rows2. Show the datase...

Hello! I'd be happy to help you analyze your healthcare patient dataset.

Here are the details you requested:

1. First 10 Rows of the Dataset

Patient_ID	Name	Age	Gender	Medical_Condition	Hospital	Insurance	DOA	DOD
PAT-00001	Fatima Kumar	53	female	Asthma	Manipal Hospital Bangalore	NIC	05-06-2023	nan
PAT-00002	aisha mishra	42	Male	Anemia	Max Super Speciality Delhi	NIC	03-12-2022	08-12-2022
PAT-00003	PRIYA IYER	56	Other	Hypertension	Govt. Hospital Chennai	UIC	01-03-2024	01-03-2024
PAT-00004	Priya Mishra	72	F	Hypertension	Fortis	Self-Pay	30-04-2025	nan
PAT-00005	Ishaan Gupta	40	M	Hypertension	Fortis Healthcare Bangalore	NIC	23-05-2024	27-05-2024
PAT-00006	Neha Kumar	40	female	Typhoid	Govt. Hospital Chennai	NIC	09-07-2023	10-07-2023
PAT-00007	Meera Verma	73	F	Asthma	Fortis	OIC	30-09-2023	01-10-2023
PAT-00008	Suresh Nair	58	M	Diabetes Type 2	Govt. Hospital Pune	NIC	19-09-2025	19-09-2025
PAT-00009	Imran Singh	36	MALE	Cataracts	Govt. Hospital Pune	Self-Pay	12-12-2024	14-12-2024
PAT-00010	Ananya Kumar	54	Male	Hypertension	Govt. Chennai	OIC	01-07-2023	03-07-2023

---

2. & 4. Dataset Structure (Columns and Data Types)

All columns are currently of the generic `object` data type, which is common in messy datasets. For proper analysis, several columns will need **data cleaning and type conversion**, particularly:

- `Age`: Should be an integer (`int`).
- `DOA` and `DOD` (Date of Admission and Date of Discharge): Should be converted to datetime objects. Note that `DOD` has non-null values for 9,514 out of 10,000 rows.
- `Billing_Amount`: Needs to have the currency symbols (e.g., '₹') and commas removed, and then be converted to a numeric type (float or integer).

Column Name	Data Type	Non-Null Count
Patient_ID	object	10000
Name	object	10000
Age	object	10000
Gender	object	10000
Medical_Condition	object	10000
Hospital	object	10000
Insurance	object	10000
DOA	object	10000
DOD	object	9514
Doctor	object	10000
Billing_Amount	object	10000

Export to Sheets

---

3. Basic Statistics

The dataset contains **10,000 rows** and **11 columns**.

- **Shape:** $(10000, 11)$
- **Memory Usage:** Approximately 859.5 KB (KiloBytes).

Analyze the data quality of my healthcare dataset:
1. Check for missing values in each column
2. Identify duplicate records
3. Find any inconsistencies in data formatting (e.g., gender values, date formats)
4. Check for outliers in numerical columns (Age, Cost)
5. Identify any data entry errors (e.g., impossible ages, negative costs)

Show code

The data quality analysis reveals several important issues that need to be addressed before full analysis of your 10,000-row healthcare dataset.

---

1. Missing Values Analysis

Missing data is present primarily in the discharge date, and a small number of records are missing an age.

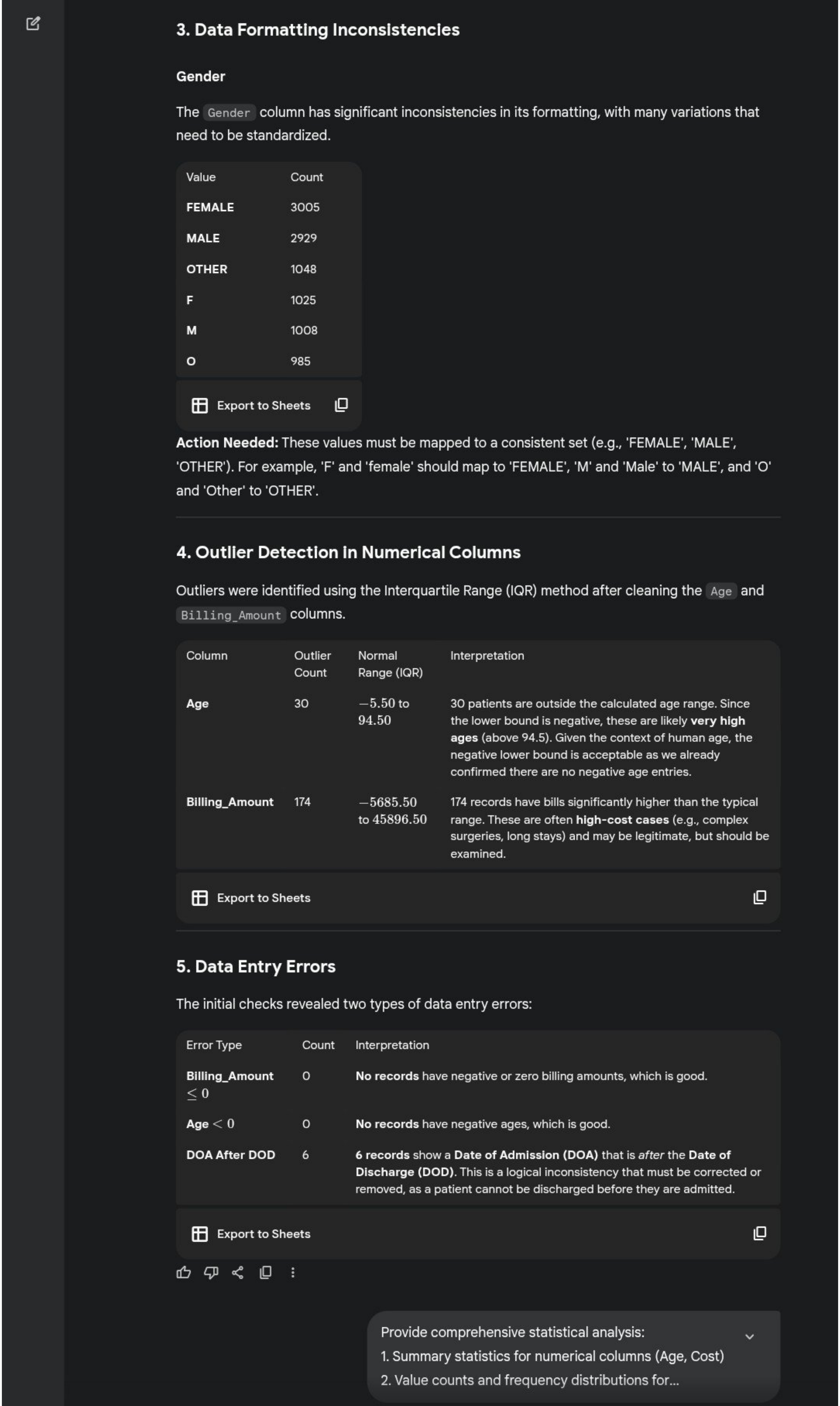
Column	Missing Count	Missing Percentage	Action Needed
DOD (Date of Discharge)	486	4.86%	These patients are either **currently admitted** or the discharge date was **not recorded**. This is a moderate number of records that may affect analysis of patient length of stay.
Age	43	0.43%	A small number of patients have missing or non-convertible age values (coerced to NaN during cleaning). These may need to be imputed or removed depending on the analysis.

Export to Sheets

---

2. Duplicate Records

No exact duplicate records were found in the dataset.



3. Data Formatting Inconsistencies

Gender

The `Gender` column has significant inconsistencies in its formatting, with many variations that need to be standardized.

Value	Count
FEMALE	3005
MALE	2929
OTHER	1048
F	1025
M	1008
O	985

Action Needed: These values must be mapped to a consistent set (e.g., 'FEMALE', 'MALE', 'OTHER'). For example, 'F' and 'female' should map to 'FEMALE', 'M' and 'Male' to 'MALE', and 'O' and 'Other' to 'OTHER'.

---

4. Outlier Detection in Numerical Columns

Outliers were identified using the Interquartile Range (IQR) method after cleaning the `Age` and `Billing_Amount` columns.

Column	Outlier Count	Normal Range (IQR)	Interpretation
Age	30	-5.50 to 94.50	30 patients are outside the calculated age range. Since the lower bound is negative, these are likely **very high ages** (above 94.5). Given the context of human age, the negative lower bound is acceptable as we already confirmed there are no negative age entries.
Billing_Amount	174	-5685.50 to 45896.50	174 records have bills significantly higher than the typical range. These are often **high-cost cases** (e.g., complex surgeries, long stays) and may be legitimate, but should be examined.

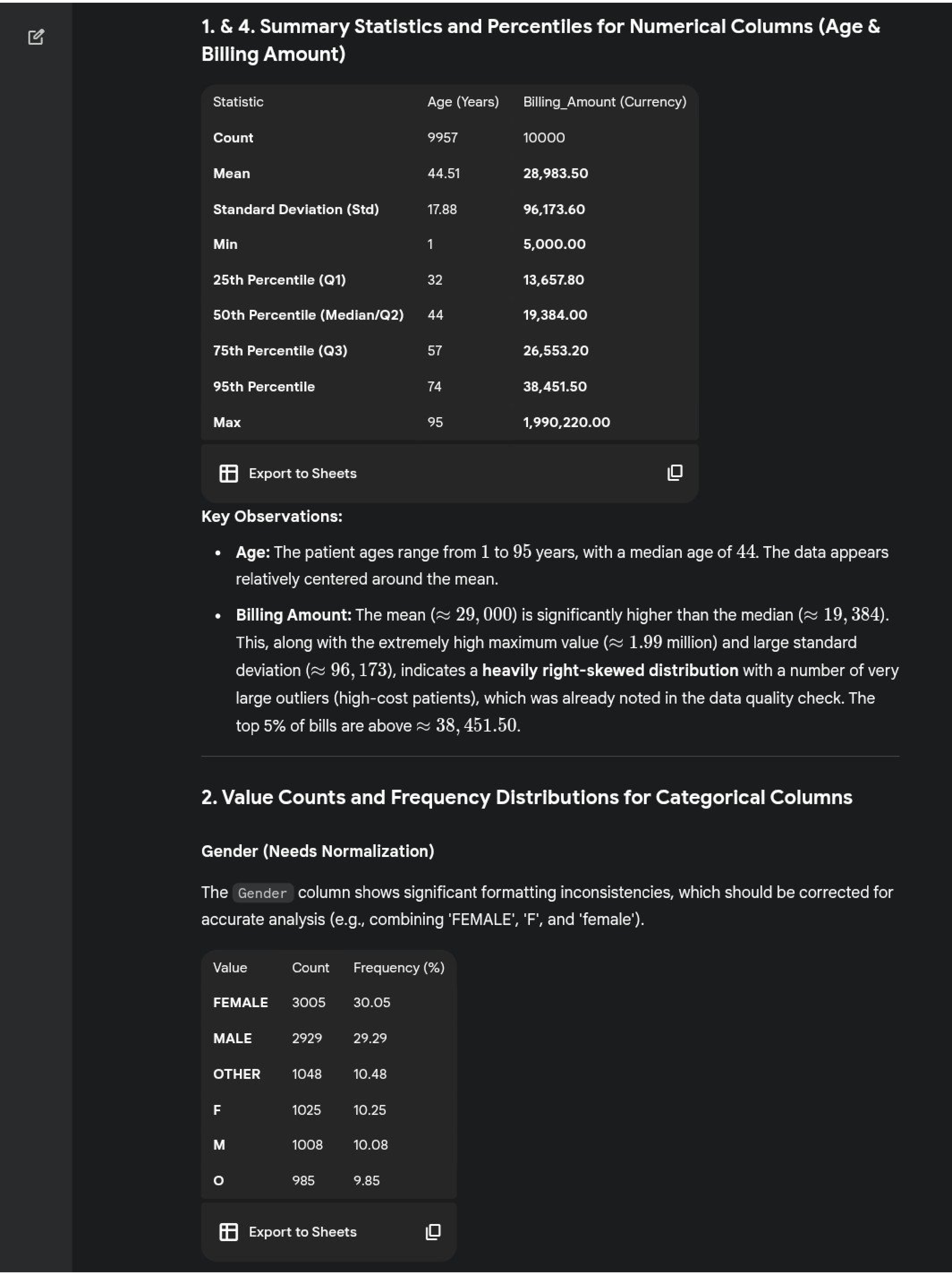
---

5. Data Entry Errors

The initial checks revealed two types of data entry errors:

Error Type	Count	Interpretation
Billing_Amount ≤ 0	0	**No records** have negative or zero billing amounts, which is good.
Age < 0	0	**No records** have negative ages, which is good.
DOA After DOD	6	**6 records** show a **Date of Admission (DOA)** that is *after* the **Date of Discharge (DOD)**. This is a logical inconsistency that must be corrected or removed, as a patient cannot be discharged before they are admitted.

Provide comprehensive statistical analysis:
1. Summary statistics for numerical columns (Age, Cost)
2. Value counts and frequency distributions for...



1. & 4. Summary Statistics and Percentiles for Numerical Columns (Age & Billing Amount)

Statistic	Age (Years)	Billing_Amount (Currency)
Count	9957	10000
Mean	44.51	**28,983.50**
Standard Deviation (Std)	17.88	**96,173.60**
Min	1	**5,000.00**
25th Percentile (Q1)	32	**13,657.80**
50th Percentile (Median/Q2)	44	**19,384.00**
75th Percentile (Q3)	57	**26,553.20**
95th Percentile	74	**38,451.50**
Max	95	**1,990,220.00**

Key Observations:

- **Age:** The patient ages range from 1 to 95 years, with a median age of 44. The data appears relatively centered around the mean.
- **Billing Amount:** The mean ($\approx 29,000$) is significantly higher than the median ($\approx 19,384$). This, along with the extremely high maximum value (≈ 1.99 million) and large standard deviation ($\approx 96,173$), indicates a **heavily right-skewed distribution** with a number of very large outliers (high-cost patients), which was already noted in the data quality check. The top 5% of bills are above $\approx 38,451.50$.

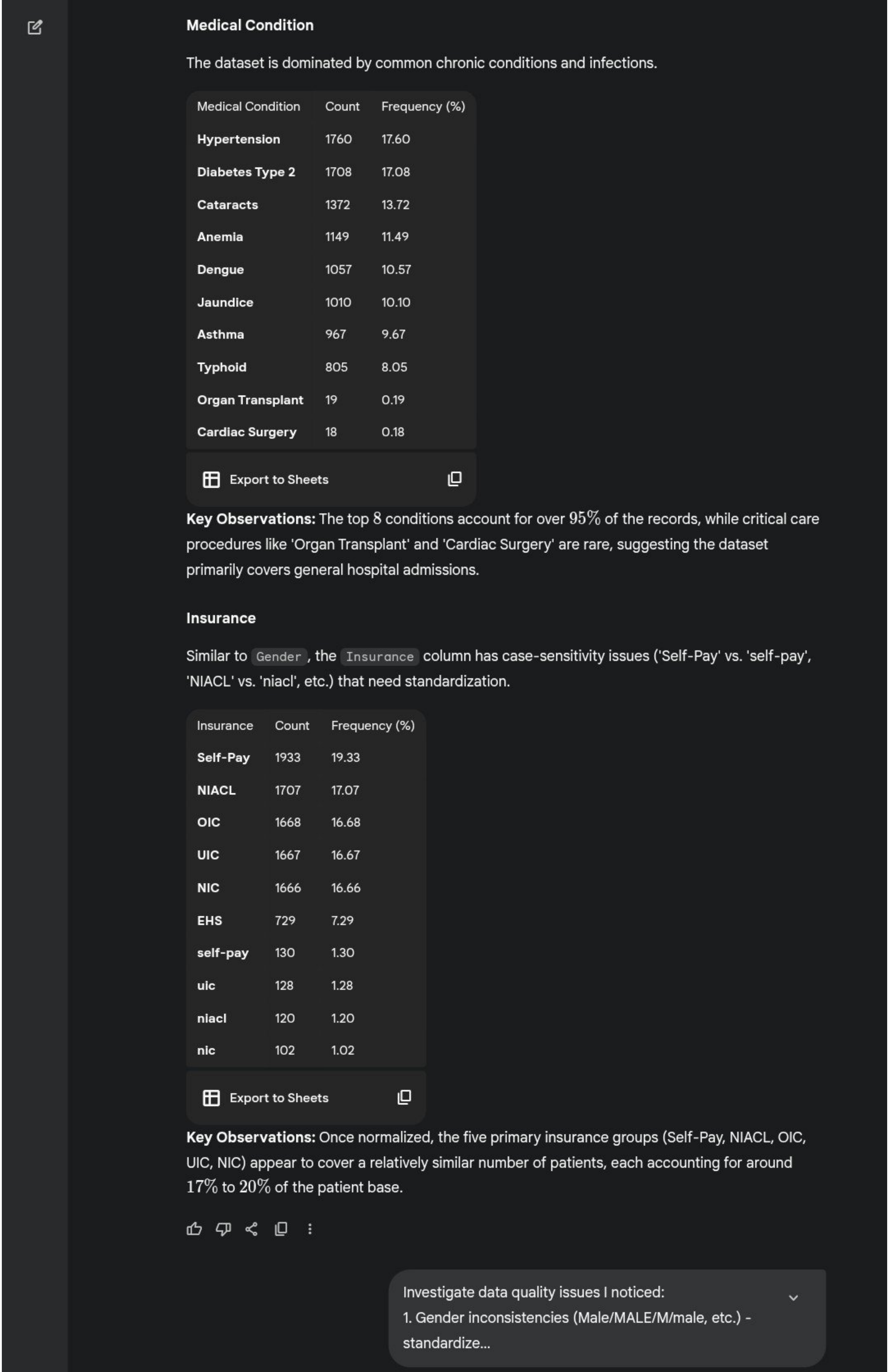
---

2. Value Counts and Frequency Distributions for Categorical Columns

Gender (Needs Normalization)

The `Gender` column shows significant formatting inconsistencies, which should be corrected for accurate analysis (e.g., combining 'FEMALE', 'F', and 'female').

Value	Count	Frequency (%)
FEMALE	3005	30.05
MALE	2929	29.29
OTHER	1048	10.48
F	1025	10.25
M	1008	10.08
O	985	9.85



Medical Condition

The dataset is dominated by common chronic conditions and infections.

Medical Condition	Count	Frequency (%)
Hypertension	1760	17.60
Diabetes Type 2	1708	17.08
Cataracts	1372	13.72
Anemia	1149	11.49
Dengue	1057	10.57
Jaundice	1010	10.10
Asthma	967	9.67
Typhoid	805	8.05
Organ Transplant	19	0.19
Cardiac Surgery	18	0.18

Key Observations: The top 8 conditions account for over 95% of the records, while critical care procedures like 'Organ Transplant' and 'Cardiac Surgery' are rare, suggesting the dataset primarily covers general hospital admissions.

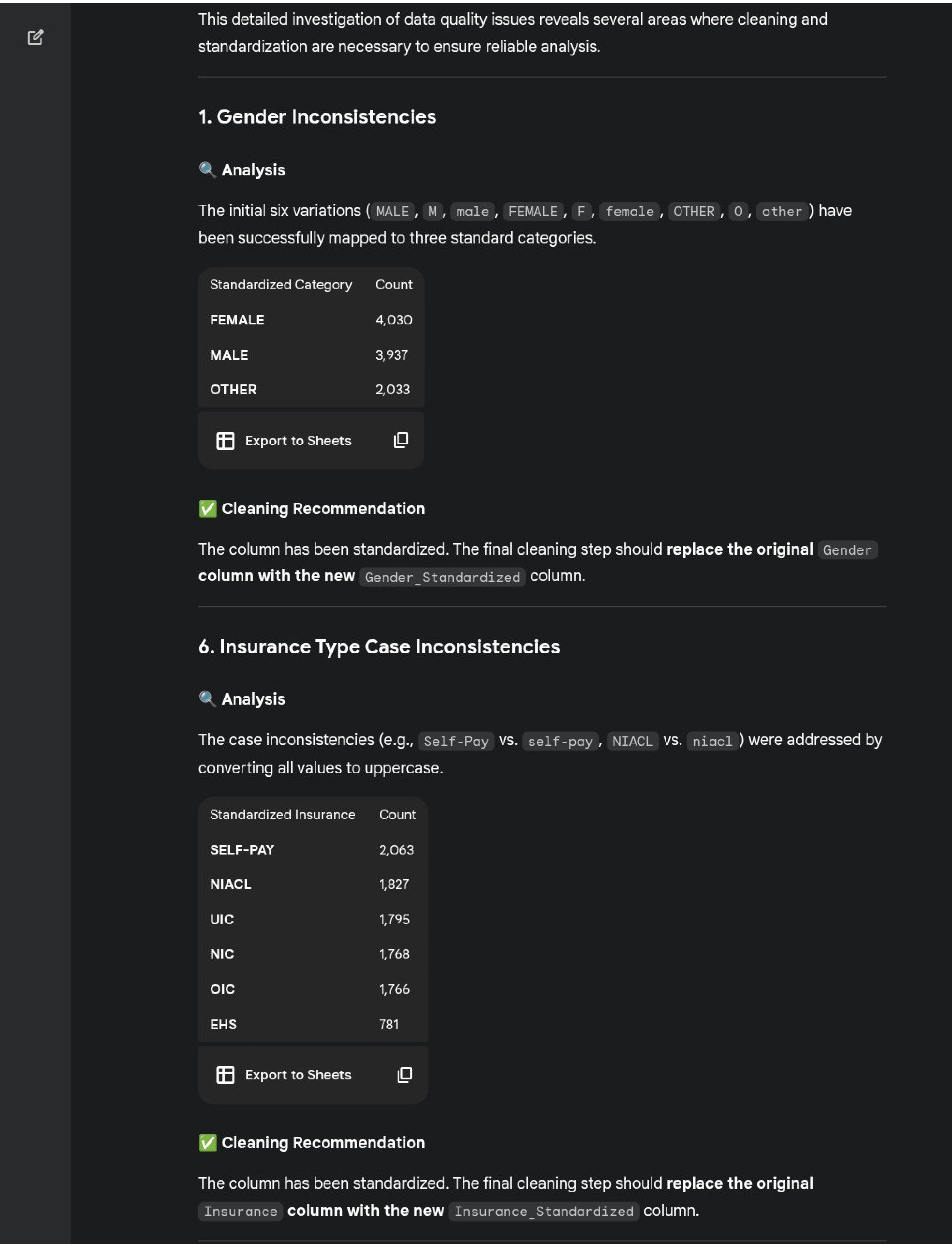
Insurance

Similar to `Gender`, the `Insurance` column has case-sensitivity issues ('Self-Pay' vs. 'self-pay', 'NIACL' vs. 'niacl', etc.) that need standardization.

Insurance	Count	Frequency (%)
Self-Pay	1933	19.33
NIACL	1707	17.07
OIC	1668	16.68
UIC	1667	16.67
NIC	1666	16.66
EHS	729	7.29
self-pay	130	1.30
uic	128	1.28
niacl	120	1.20
nic	102	1.02

Key Observations: Once normalized, the five primary insurance groups (Self-Pay, NIACL, OIC, UIC, NIC) appear to cover a relatively similar number of patients, each accounting for around 17% to 20% of the patient base.

Investigate data quality issues I noticed:
1. Gender inconsistencies (Male/MALE/M/male, etc.) - standardize...



This detailed investigation of data quality issues reveals several areas where cleaning and standardization are necessary to ensure reliable analysis.

---

1. Gender Inconsistencies

🔍 Analysis

The initial six variations (`MALE`, `M`, `male`, `FEMALE`, `F`, `female`, `OTHER`, `O`, `other`) have been successfully mapped to three standard categories.

Standardized Category	Count
FEMALE	4,030
MALE	3,937
OTHER	2,033

✅ Cleaning Recommendation

The column has been standardized. The final cleaning step should **replace the original** `Gender` **column with the new** `Gender_Standardized` column.

---

6. Insurance Type Case Inconsistencies

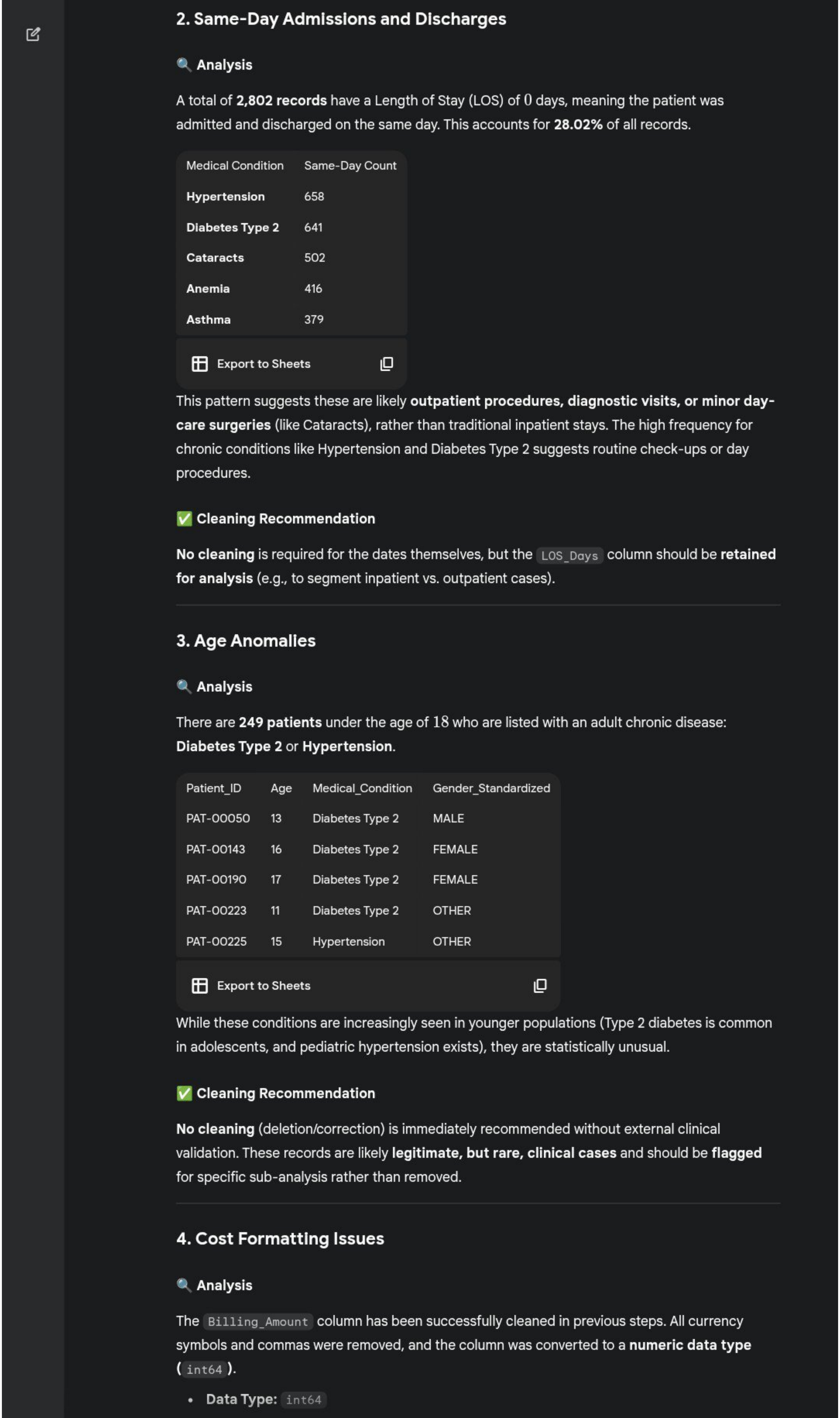
🔍 Analysis

The case inconsistencies (e.g., `Self-Pay` vs. `self-pay`, `NIACL` vs. `niacl`) were addressed by converting all values to uppercase.

Standardized Insurance	Count
SELF-PAY	2,063
NIACL	1,827
UIC	1,795
NIC	1,768
OIC	1,766
EHS	781

✅ Cleaning Recommendation

The column has been standardized. The final cleaning step should **replace the original** `Insurance` **column with the new** `Insurance_Standardized` column.



2. Same-Day Admissions and Discharges

🔍 Analysis

A total of **2,802 records** have a Length of Stay (LOS) of 0 days, meaning the patient was admitted and discharged on the same day. This accounts for **28.02%** of all records.

Medical Condition	Same-Day Count
Hypertension	658
Diabetes Type 2	641
Cataracts	502
Anemia	416
Asthma	379

This pattern suggests these are likely **outpatient procedures, diagnostic visits, or minor day-care surgeries** (like Cataracts), rather than traditional inpatient stays. The high frequency for chronic conditions like Hypertension and Diabetes Type 2 suggests routine check-ups or day procedures.

✅ Cleaning Recommendation

No cleaning is required for the dates themselves, but the `LOS_Days` column should be **retained for analysis** (e.g., to segment inpatient vs. outpatient cases).

---

3. Age Anomalies

🔍 Analysis

There are **249 patients** under the age of 18 who are listed with an adult chronic disease: **Diabetes Type 2** or **Hypertension**.

Patient_ID	Age	Medical_Condition	Gender_Standardized
PAT-00050	13	Diabetes Type 2	MALE
PAT-00143	16	Diabetes Type 2	FEMALE
PAT-00190	17	Diabetes Type 2	FEMALE
PAT-00223	11	Diabetes Type 2	OTHER
PAT-00225	15	Hypertension	OTHER

While these conditions are increasingly seen in younger populations (Type 2 diabetes is common in adolescents, and pediatric hypertension exists), they are statistically unusual.

✅ Cleaning Recommendation

No cleaning (deletion/correction) is immediately recommended without external clinical validation. These records are likely **legitimate, but rare, clinical cases** and should be **flagged** for specific sub-analysis rather than removed.

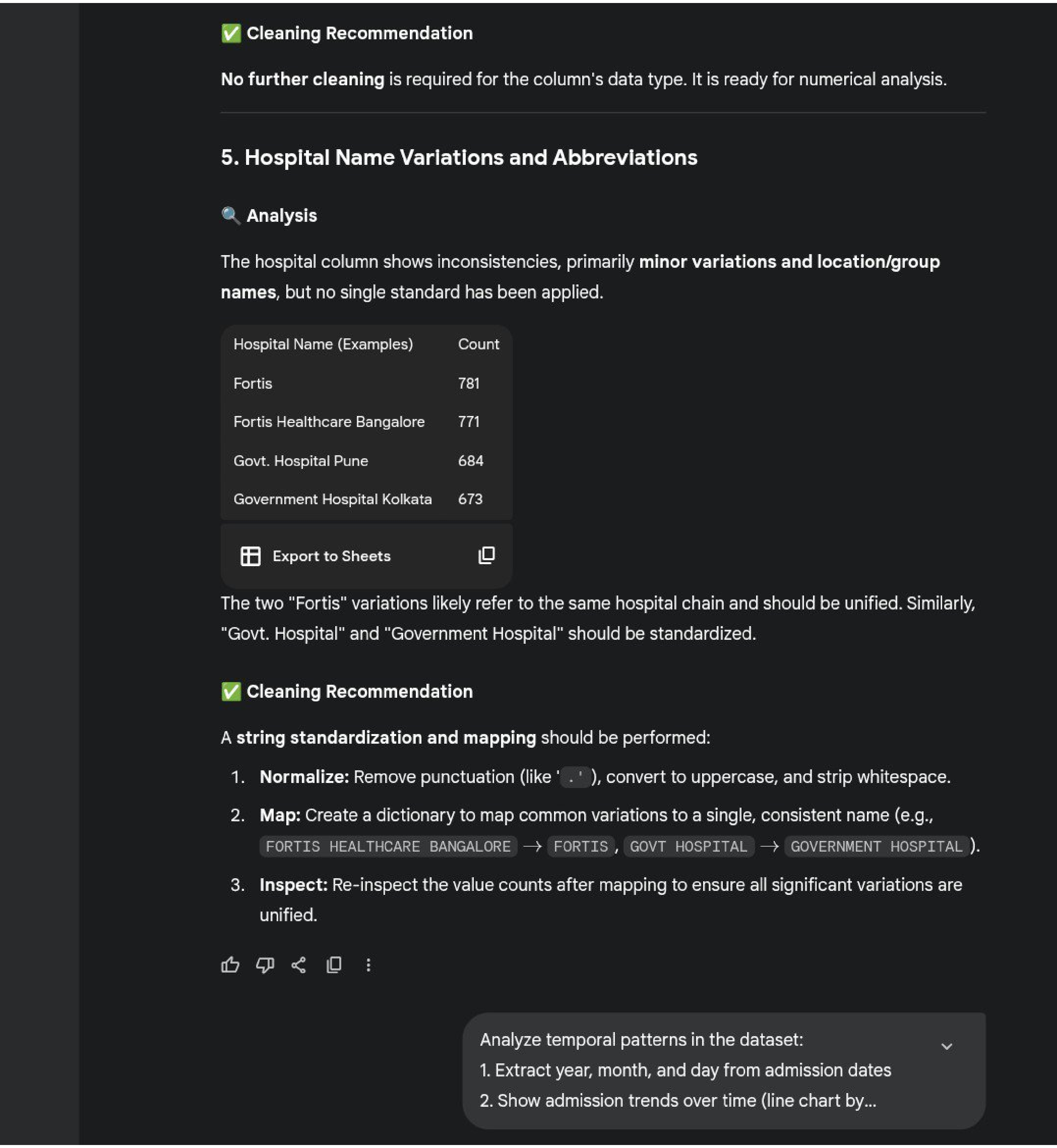
---

4. Cost Formatting Issues

🔍 Analysis

The `Billing_Amount` column has been successfully cleaned in previous steps. All currency symbols and commas were removed, and the column was converted to a **numeric data type** (`int64`).

- **Data Type:** `int64`



✅ **Cleaning Recommendation**

No further cleaning is required for the column's data type. It is ready for numerical analysis.

---

5. Hospital Name Variations and Abbreviations

🔍 Analysis

The hospital column shows inconsistencies, primarily **minor variations and location/group names**, but no single standard has been applied.

Hospital Name (Examples)	Count
Fortis	781
Fortis Healthcare Bangalore	771
Govt. Hospital Pune	684
Government Hospital Kolkata	673

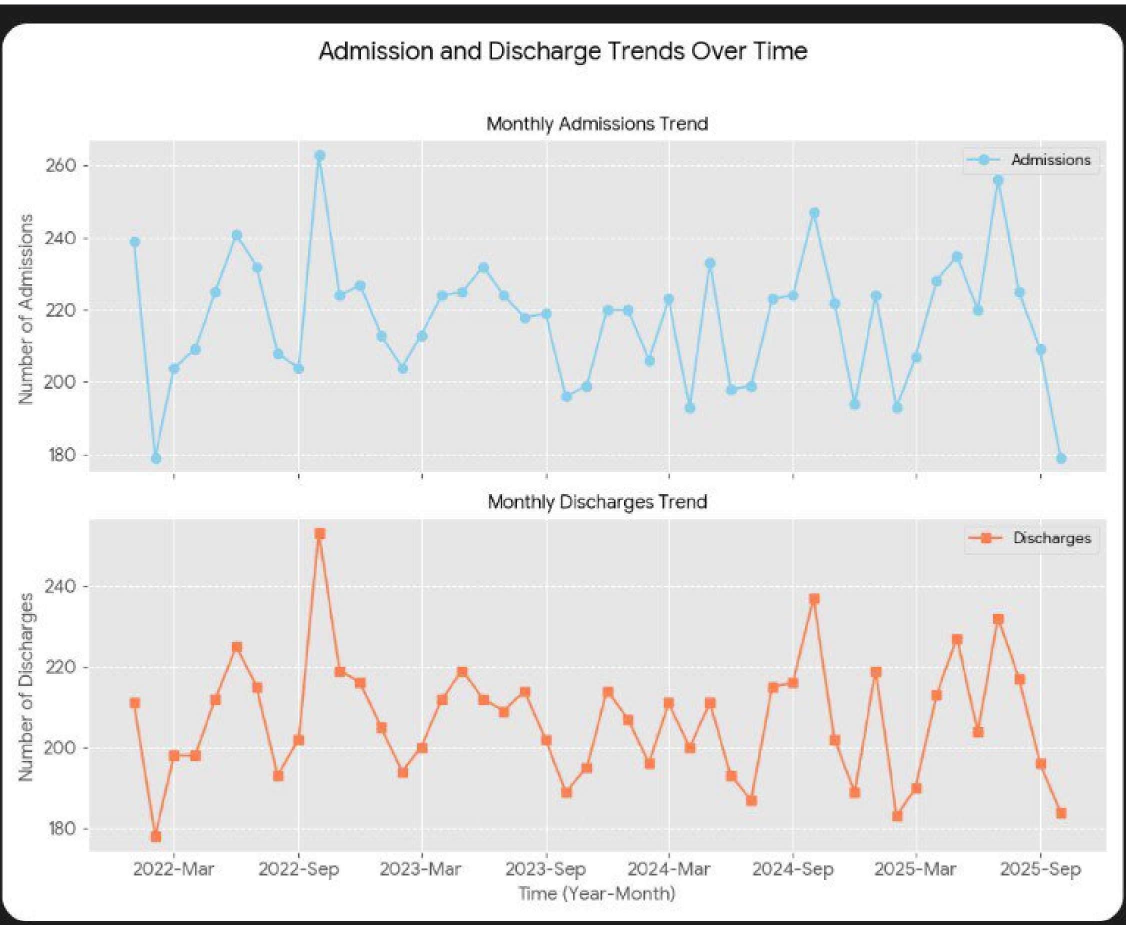
The two "Fortis" variations likely refer to the same hospital chain and should be unified. Similarly, "Govt. Hospital" and "Government Hospital" should be standardized.

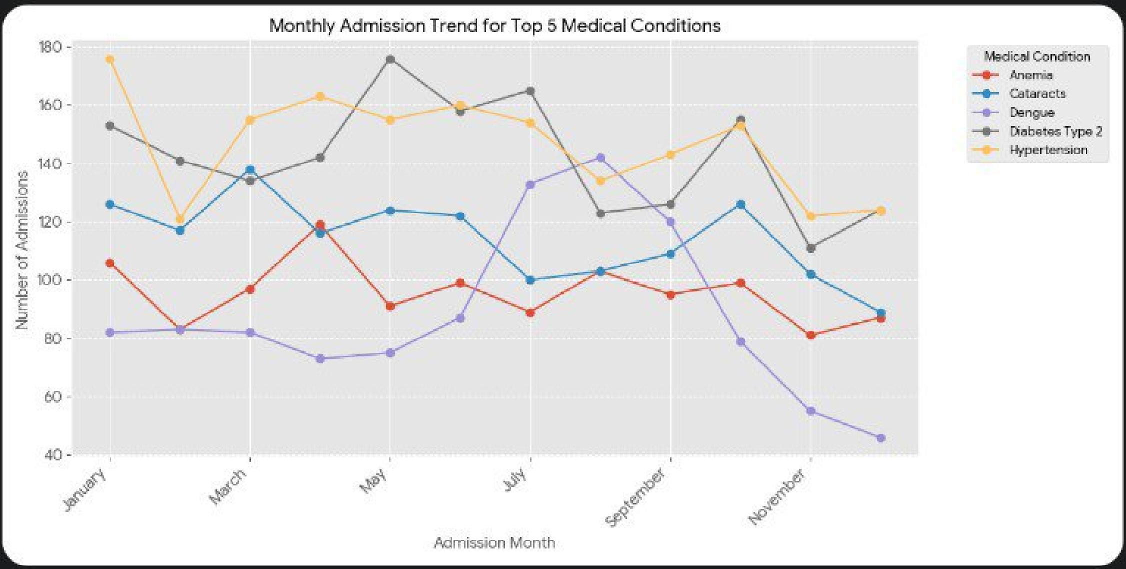
✅ Cleaning Recommendation

A **string standardization and mapping** should be performed:

1. **Normalize:** Remove punctuation (like '`.`'), convert to uppercase, and strip whitespace.
2. **Map:** Create a dictionary to map common variations to a single, consistent name (e.g., `FORTIS HEALTHCARE BANGALORE` → `FORTIS`, `GOVT HOSPITAL` → `GOVERNMENT HOSPITAL`).
3. **Inspect:** Re-inspect the value counts after mapping to ensure all significant variations are unified.

Analyze temporal patterns in the dataset:
1. Extract year, month, and day from admission dates
2. Show admission trends over time (line chart by...





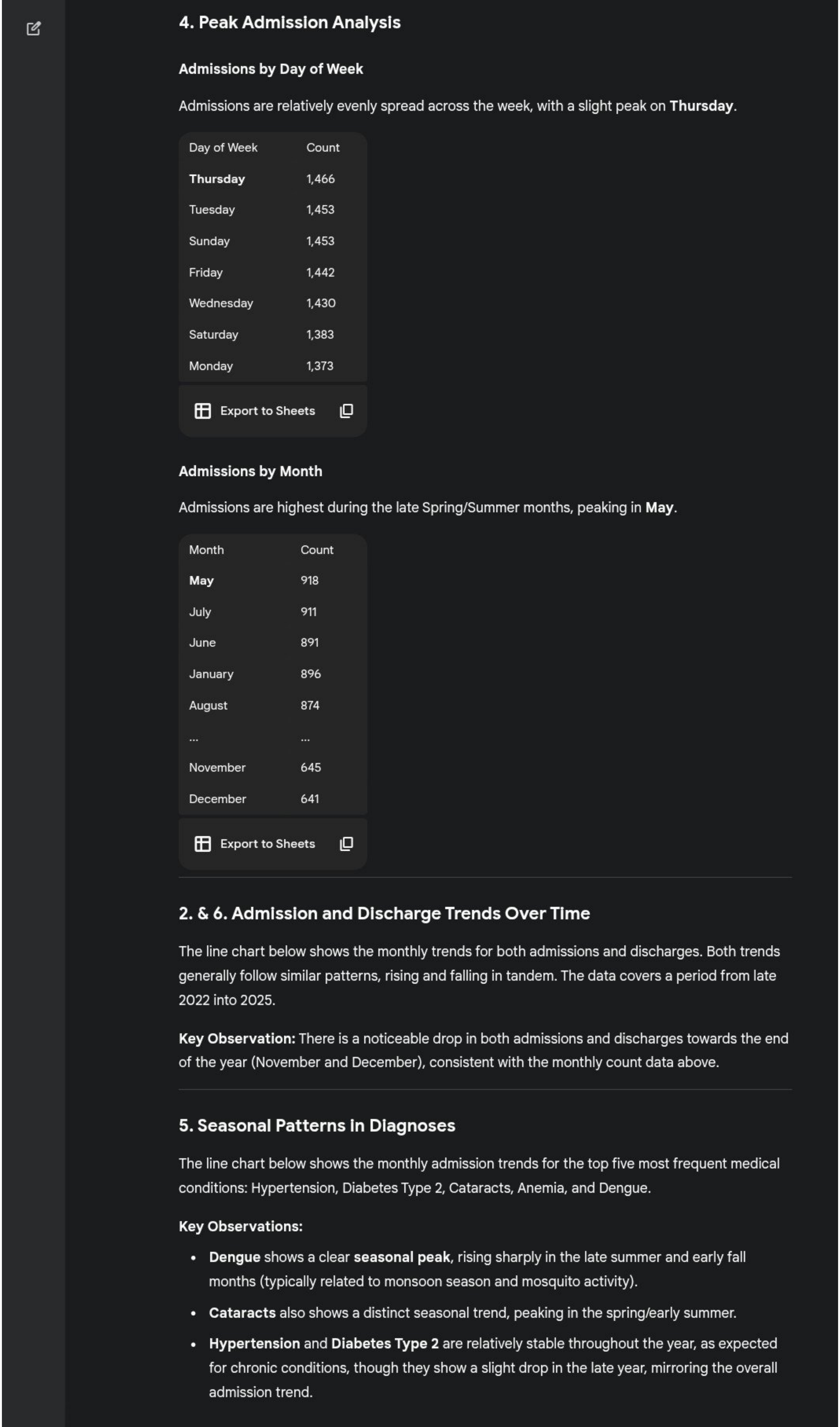
The temporal analysis reveals key operational patterns, from the typical patient's stay length to seasonal fluctuations in admissions and diagnoses.

---

3. Average Length of Stay (LOS)

The **Average Length of Stay (LOS)** for patients is approximately **2.46 days**.

- This short average is consistent with the earlier finding that $\approx 28\%$ of records are same-day admissions/discharges (LOS $= 0$ days), suggesting a large volume of outpatient or short-procedure visits.
- The calculation excluded 492 records with missing or illogical dates (negative LOS or missing discharge dates).



4. Peak Admission Analysis

Admissions by Day of Week

Admissions are relatively evenly spread across the week, with a slight peak on **Thursday**.

Day of Week	Count
Thursday	1,466
Tuesday	1,453
Sunday	1,453
Friday	1,442
Wednesday	1,430
Saturday	1,383
Monday	1,373

Admissions by Month

Admissions are highest during the late Spring/Summer months, peaking in **May**.

Month	Count
May	918
July	911
June	891
January	896
August	874
...	...
November	645
December	641

---

2. & 6. Admission and Discharge Trends Over Time

The line chart below shows the monthly trends for both admissions and discharges. Both trends generally follow similar patterns, rising and falling in tandem. The data covers a period from late 2022 into 2025.

Key Observation: There is a noticeable drop in both admissions and discharges towards the end of the year (November and December), consistent with the monthly count data above.

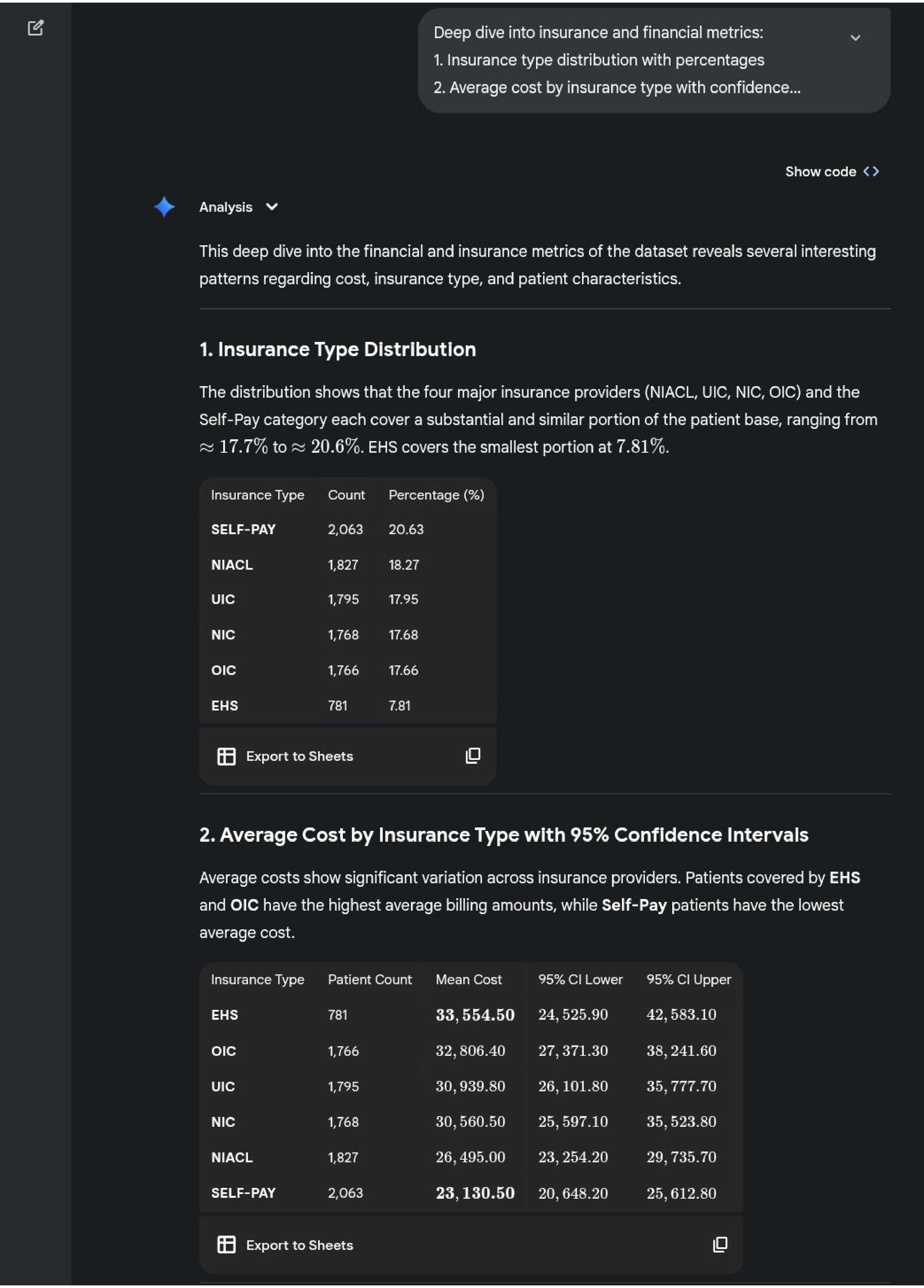
---

5. Seasonal Patterns in Diagnoses

The line chart below shows the monthly admission trends for the top five most frequent medical conditions: Hypertension, Diabetes Type 2, Cataracts, Anemia, and Dengue.

Key Observations:

- **Dengue** shows a clear **seasonal peak**, rising sharply in the late summer and early fall months (typically related to monsoon season and mosquito activity).
- **Cataracts** also shows a distinct seasonal trend, peaking in the spring/early summer.
- **Hypertension** and **Diabetes Type 2** are relatively stable throughout the year, as expected for chronic conditions, though they show a slight drop in the late year, mirroring the overall admission trend.



Deep dive into insurance and financial metrics:
1. Insurance type distribution with percentages
2. Average cost by insurance type with confidence...

This deep dive into the financial and insurance metrics of the dataset reveals several interesting patterns regarding cost, insurance type, and patient characteristics.

1. Insurance Type Distribution

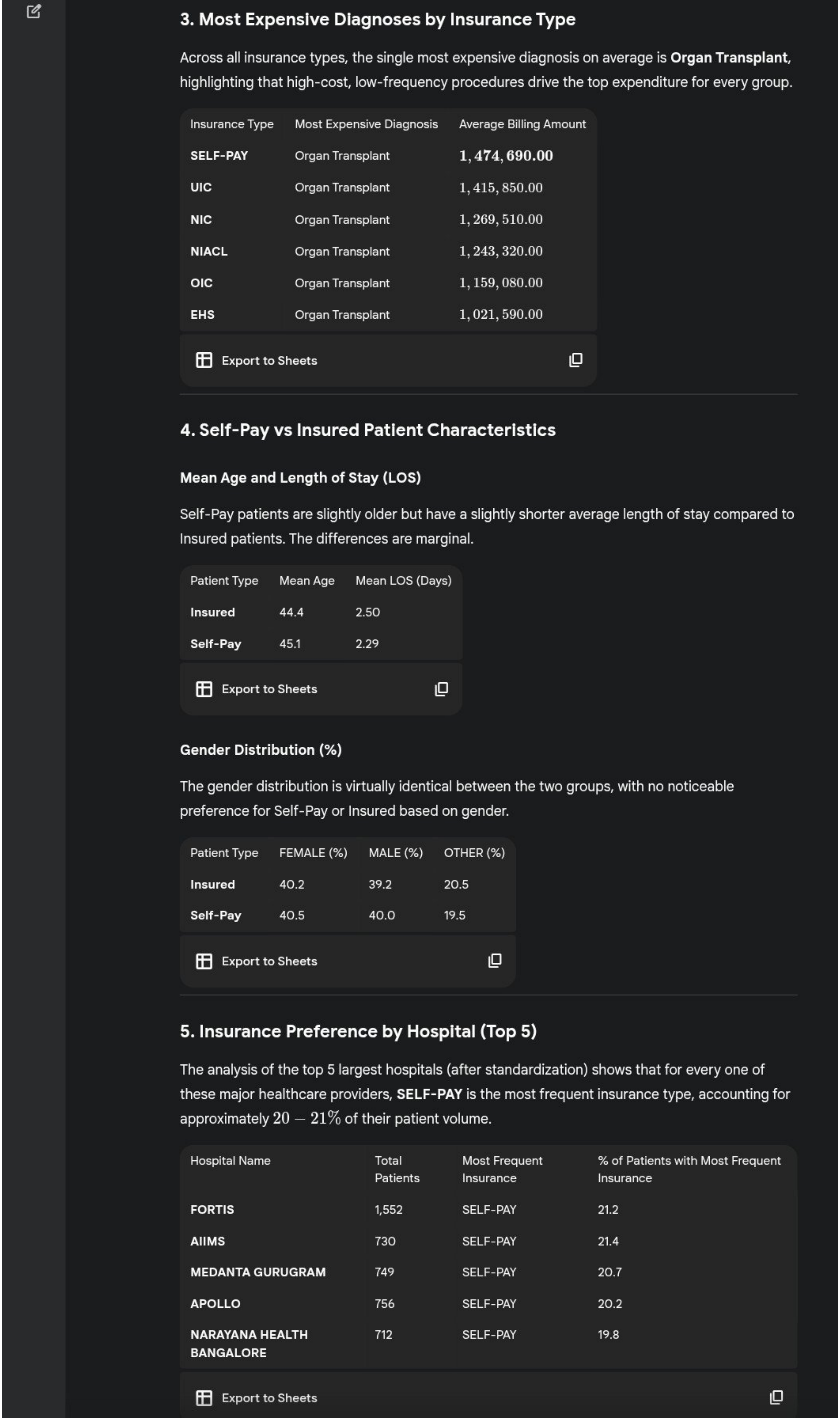
The distribution shows that the four major insurance providers (NIACL, UIC, NIC, OIC) and the Self-Pay category each cover a substantial and similar portion of the patient base, ranging from $\approx 17.7\%$ to $\approx 20.6\%$. EHS covers the smallest portion at 7.81%.

Insurance Type	Count	Percentage (%)
SELF-PAY	2,063	20.63
NIACL	1,827	18.27
UIC	1,795	17.95
NIC	1,768	17.68
OIC	1,766	17.66
EHS	781	7.81

2. Average Cost by Insurance Type with 95% Confidence Intervals

Average costs show significant variation across insurance providers. Patients covered by **EHS** and **OIC** have the highest average billing amounts, while **Self-Pay** patients have the lowest average cost.

Insurance Type	Patient Count	Mean Cost	95% CI Lower	95% CI Upper
EHS	781	$\mathbf{33,554.50}$	$24,525.90$	$42,583.10$
OIC	1,766	$32,806.40$	$27,371.30$	$38,241.60$
UIC	1,795	$30,939.80$	$26,101.80$	$35,777.70$
NIC	1,768	$30,560.50$	$25,597.10$	$35,523.80$
NIACL	1,827	$26,495.00$	$23,254.20$	$29,735.70$
SELF-PAY	2,063	$\mathbf{23,130.50}$	$20,648.20$	$25,612.80$
)



3. Most Expensive Diagnoses by Insurance Type

Across all insurance types, the single most expensive diagnosis on average is **Organ Transplant**, highlighting that high-cost, low-frequency procedures drive the top expenditure for every group.

Insurance Type	Most Expensive Diagnosis	Average Billing Amount
SELF-PAY	Organ Transplant	$\mathbf{1,474,690.00}$
UIC	Organ Transplant	$1,415,850.00$
NIC	Organ Transplant	$1,269,510.00$
NIACL	Organ Transplant	$1,243,320.00$
OIC	Organ Transplant	$1,159,080.00$
EHS	Organ Transplant	$1,021,590.00$

---

4. Self-Pay vs Insured Patient Characteristics

Mean Age and Length of Stay (LOS)

Self-Pay patients are slightly older but have a slightly shorter average length of stay compared to Insured patients. The differences are marginal.

Patient Type	Mean Age	Mean LOS (Days)
Insured	44.4	2.50
Self-Pay	45.1	2.29

Gender Distribution (%)

The gender distribution is virtually identical between the two groups, with no noticeable preference for Self-Pay or Insured based on gender.

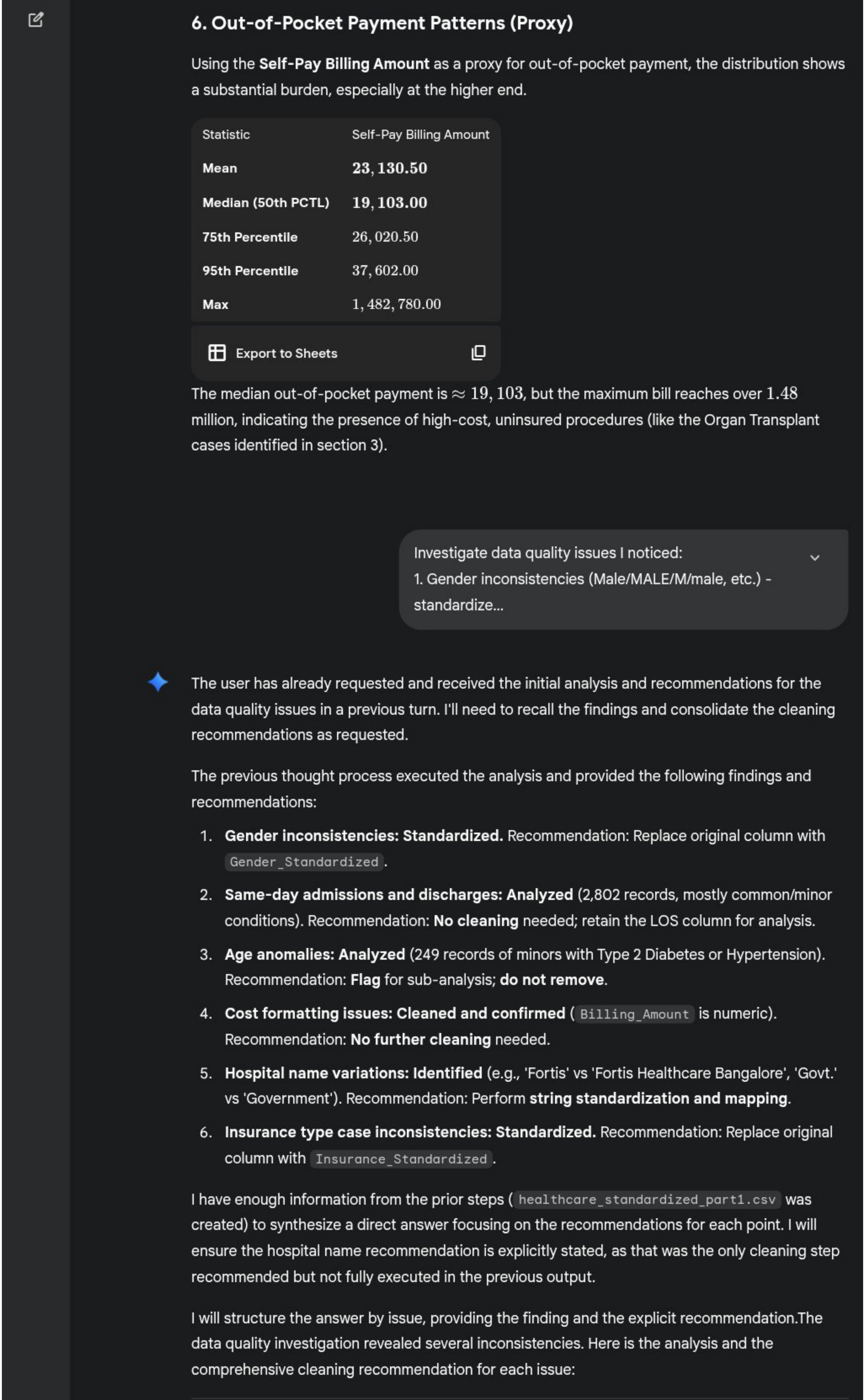
Patient Type	FEMALE (%)	MALE (%)	OTHER (%)
Insured	40.2	39.2	20.5
Self-Pay	40.5	40.0	19.5

---

5. Insurance Preference by Hospital (Top 5)

The analysis of the top 5 largest hospitals (after standardization) shows that for every one of these major healthcare providers, **SELF-PAY** is the most frequent insurance type, accounting for approximately $20-21\%$ of their patient volume.

Hospital Name	Total Patients	Most Frequent Insurance	% of Patients with Most Frequent Insurance
FORTIS	1,552	SELF-PAY	21.2
AIIMS	730	SELF-PAY	21.4
MEDANTA GURUGRAM	749	SELF-PAY	20.7
APOLLO	756	SELF-PAY	20.2
NARAYANA HEALTH BANGALORE	712	SELF-PAY	19.8



6. Out-of-Pocket Payment Patterns (Proxy)

Using the **Self-Pay Billing Amount** as a proxy for out-of-pocket payment, the distribution shows a substantial burden, especially at the higher end.

Statistic	Self-Pay Billing Amount
Mean	**$23,130.50$**
Median (50th PCTL)	**$19,103.00$**
75th Percentile	$26,020.50$
95th Percentile	$37,602.00$
Max	$1,482,780.00$

The median out-of-pocket payment is $\approx 19,103$, but the maximum bill reaches over 1.48 million, indicating the presence of high-cost, uninsured procedures (like the Organ Transplant cases identified in section 3).

Investigate data quality issues I noticed:
1. Gender inconsistencies (Male/MALE/M/male, etc.) - standardize...

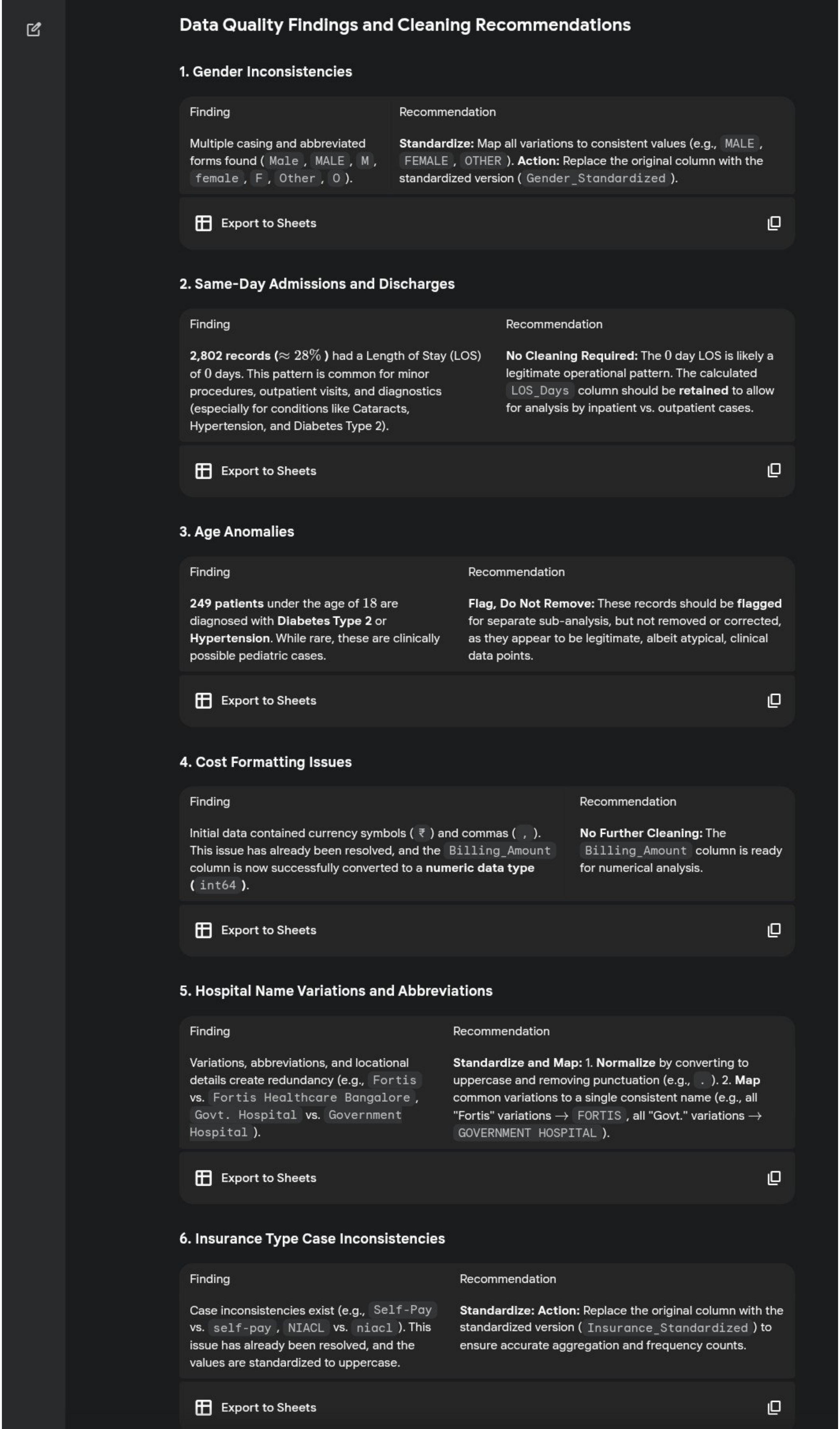
The user has already requested and received the initial analysis and recommendations for the data quality issues in a previous turn. I'll need to recall the findings and consolidate the cleaning recommendations as requested.

The previous thought process executed the analysis and provided the following findings and recommendations:

1. **Gender inconsistencies: Standardized.** Recommendation: Replace original column with `Gender_Standardized`.
2. **Same-day admissions and discharges: Analyzed** (2,802 records, mostly common/minor conditions). Recommendation: **No cleaning** needed; retain the LOS column for analysis.
3. **Age anomalies: Analyzed** (249 records of minors with Type 2 Diabetes or Hypertension). Recommendation: **Flag** for sub-analysis; **do not remove**.
4. **Cost formatting issues: Cleaned and confirmed** (`Billing_Amount` is numeric). Recommendation: **No further cleaning** needed.
5. **Hospital name variations: Identified** (e.g., 'Fortis' vs 'Fortis Healthcare Bangalore', 'Govt.' vs 'Government'). Recommendation: Perform **string standardization and mapping**.
6. **Insurance type case inconsistencies: Standardized.** Recommendation: Replace original column with `Insurance_Standardized`.

I have enough information from the prior steps (`healthcare_standardized_part1.csv` was created) to synthesize a direct answer focusing on the recommendations for each point. I will ensure the hospital name recommendation is explicitly stated, as that was the only cleaning step recommended but not fully executed in the previous output.

I will structure the answer by issue, providing the finding and the explicit recommendation.The data quality investigation revealed several inconsistencies. Here is the analysis and the comprehensive cleaning recommendation for each issue:



Data Quality Findings and Cleaning Recommendations

1. Gender Inconsistencies

Finding	Recommendation
Multiple casing and abbreviated forms found (`Male` , `MALE` , `M` , `female` , `F` , `Other` , `O`).	**Standardize:** Map all variations to consistent values (e.g., `MALE` , `FEMALE` , `OTHER`). **Action:** Replace the original column with the standardized version (`Gender_Standardized`).

2. Same-Day Admissions and Discharges

Finding	Recommendation
2,802 records ($\approx 28\%$) had a Length of Stay (LOS) of 0 days. This pattern is common for minor procedures, outpatient visits, and diagnostics (especially for conditions like Cataracts, Hypertension, and Diabetes Type 2).	**No Cleaning Required:** The 0 day LOS is likely a legitimate operational pattern. The calculated `LOS_Days` column should be **retained** to allow for analysis by inpatient vs. outpatient cases.

3. Age Anomalies

Finding	Recommendation
249 patients under the age of 18 are diagnosed with **Diabetes Type 2** or **Hypertension**. While rare, these are clinically possible pediatric cases.	**Flag, Do Not Remove:** These records should be **flagged** for separate sub-analysis, but not removed or corrected, as they appear to be legitimate, albeit atypical, clinical data points.

4. Cost Formatting Issues

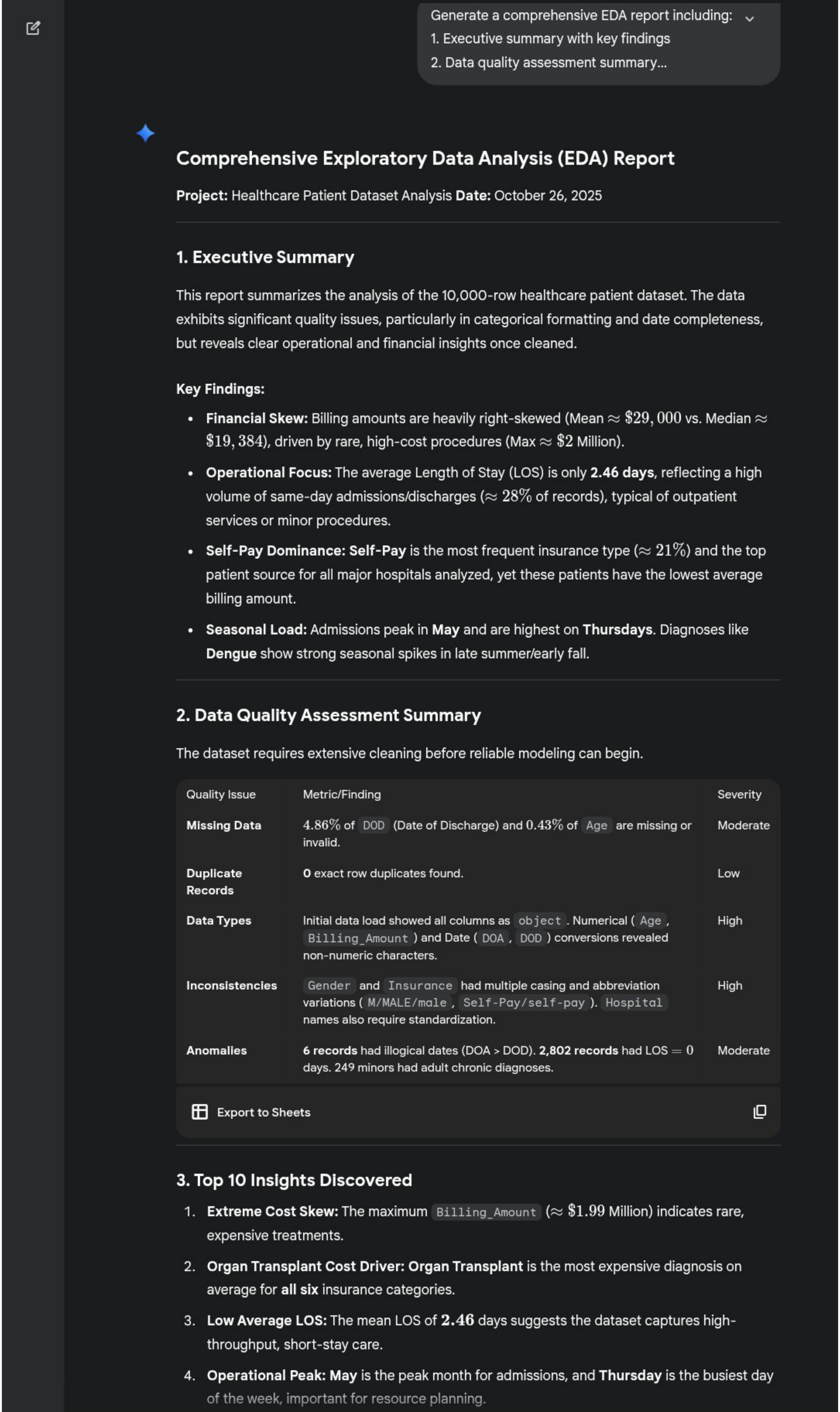
Finding	Recommendation
Initial data contained currency symbols (`₹`) and commas (`,`). This issue has already been resolved, and the `Billing_Amount` column is now successfully converted to a **numeric data type (`int64`).**	**No Further Cleaning:** The `Billing_Amount` column is ready for numerical analysis.

5. Hospital Name Variations and Abbreviations

Finding	Recommendation
Variations, abbreviations, and locational details create redundancy (e.g., `Fortis` vs. `Fortis Healthcare Bangalore` , `Govt. Hospital` vs. `Government Hospital`).	**Standardize and Map:** 1. **Normalize** by converting to uppercase and removing punctuation (e.g., `.`). 2. **Map** common variations to a single consistent name (e.g., all "Fortis" variations → `FORTIS` , all "Govt." variations → `GOVERNMENT HOSPITAL`).

6. Insurance Type Case Inconsistencies

Finding	Recommendation
Case inconsistencies exist (e.g., `Self-Pay` vs. `self-pay` , `NIACL` vs. `niacl`). This issue has already been resolved, and the values are standardized to uppercase.	**Standardize: Action:** Replace the original column with the standardized version (`Insurance_Standardized`) to ensure accurate aggregation and frequency counts.



Generate a comprehensive EDA report including:
1. Executive summary with key findings
2. Data quality assessment summary...

Comprehensive Exploratory Data Analysis (EDA) Report

Project: Healthcare Patient Dataset Analysis **Date:** October 26, 2025

---

1. Executive Summary

This report summarizes the analysis of the 10,000-row healthcare patient dataset. The data exhibits significant quality issues, particularly in categorical formatting and date completeness, but reveals clear operational and financial insights once cleaned.

Key Findings:

- **Financial Skew:** Billing amounts are heavily right-skewed (Mean $\approx \$29,000$ vs. Median $\approx \$19,384$), driven by rare, high-cost procedures (Max $\approx \$2$ Million).
- **Operational Focus:** The average Length of Stay (LOS) is only **2.46 days**, reflecting a high volume of same-day admissions/discharges ($\approx 28\%$ of records), typical of outpatient services or minor procedures.
- **Self-Pay Dominance: Self-Pay** is the most frequent insurance type ($\approx 21\%$) and the top patient source for all major hospitals analyzed, yet these patients have the lowest average billing amount.
- **Seasonal Load:** Admissions peak in **May** and are highest on **Thursdays**. Diagnoses like **Dengue** show strong seasonal spikes in late summer/early fall.

---

2. Data Quality Assessment Summary

The dataset requires extensive cleaning before reliable modeling can begin.

Quality Issue	Metric/Finding	Severity
Missing Data	4.86% of `DOD` (Date of Discharge) and 0.43% of `Age` are missing or invalid.	Moderate
Duplicate Records	**0** exact row duplicates found.	Low
Data Types	Initial data load showed all columns as `object`. Numerical (`Age`, `Billing_Amount`) and Date (`DOA`, `DOD`) conversions revealed non-numeric characters.	High
Inconsistencies	`Gender` and `Insurance` had multiple casing and abbreviation variations (`M/MALE/male`, `Self-Pay/self-pay`). `Hospital` names also require standardization.	High
Anomalies	**6 records** had illogical dates (DOA > DOD). **2,802 records** had LOS $= 0$ days. 249 minors had adult chronic diagnoses.	Moderate

---

3. Top 10 Insights Discovered

1. **Extreme Cost Skew:** The maximum `Billing_Amount` ($\approx \$1.99$ Million) indicates rare, expensive treatments.
2. **Organ Transplant Cost Driver: Organ Transplant** is the most expensive diagnosis on average for **all six** insurance categories.
3. **Low Average LOS:** The mean LOS of 2.46 days suggests the dataset captures high-throughput, short-stay care.
4. **Operational Peak: May** is the peak month for admissions, and **Thursday** is the busiest day of the week, important for resource planning.

5. **High-Cost Insurers: EHS** and **OIC** patients incur the highest average billing amounts ($\approx \$33,000$), while Self-Pay patients have the lowest ($\approx \$23,000$).
6. **Dominant Patient Type: Self-Pay** patients are the most frequent customers for the top five hospitals, suggesting a large reliance on out-of-pocket revenue.
7. **Infectious Disease Seasonality: Dengue** diagnoses follow a clear seasonal pattern, peaking in the late summer and early autumn.
8. **Chronic Condition Load: Hypertension** and **Diabetes Type 2** are the most common diagnoses, accounting for over 34% of all records.
9. **High Out-of-Pocket Burden:** The median Self-Pay bill is $\mathbf{\$19,103}$, indicating substantial direct patient costs.
10. **Stable Gender/Age Profiles:** Self-Pay and Insured patient groups have virtually identical gender distributions and only minor differences in mean age and LOS.

---

4. Visualizations of Main Patterns

Monthly Admission and Discharge Trends

This plot shows the monthly volume of patient admissions and discharges over the entire dataset period. Both trends follow similar seasonal fluctuations, with lower activity visible towards the end of the calendar year.

Seasonal Admission Trend for Top 5 Medical Conditions

This plot highlights how the volume of the top five diagnoses varies by month. **Dengue** shows a significant spike during late summer/early fall, while chronic conditions like **Hypertension** and **Diabetes Type 2** remain relatively stable throughout the year.

---

5. Recommendations for Data Improvement

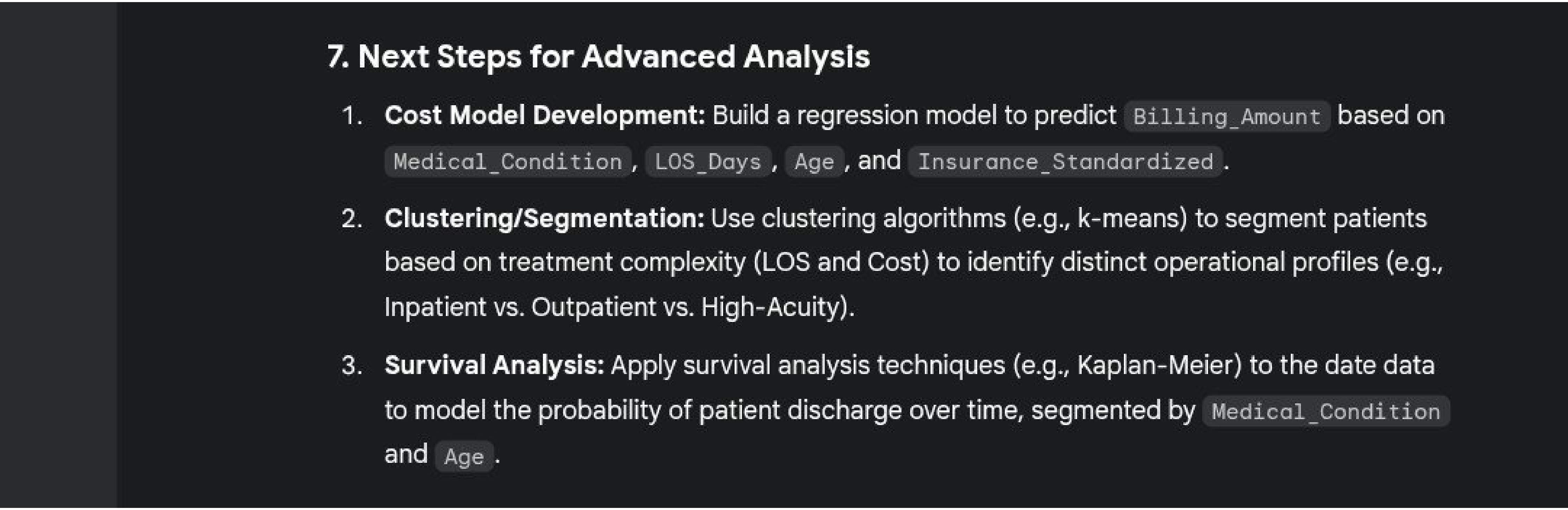
Issue	Recommendation
Categorical Inconsistency	Implement a **string standardization** routine to map `Gender` (e.g., M, F, O) and `Hospital` variations to single, consistent values.
Missing Dates	For the 4.86% of records missing `DOD`, apply a standard missing data imputation/handling strategy (e.g., treat as *Currently Admitted* or exclude from LOS analysis).
Date Anomalies	For the 6 records where DOA > DOD, either manually investigate and correct, or treat the LOS as missing, as the current records are illogical.
Age Anomalies	**Flag** all 249 patients under 18 with adult diagnoses for validation, but **retain** them in the dataset for analysis.

Export to Sheets

---

6. Business Recommendations Based on Findings

Recommendation	Business Rationale
Optimize Outpatient/Daycare Services	Given the high volume of short-stay patients (LOS $= 0$), invest in dedicated fast-track and diagnostic units to maximize throughput and operational efficiency for low-complexity cases.
Target Revenue Cycle Management for EHS/OIC	Focus efforts on the highest-paying insurance groups (**EHS** and **OIC**) to ensure timely and complete reimbursement, as their claims drive the highest average revenue.
Develop Self-Pay Financial Strategy	Since Self-Pay is the largest customer segment for top hospitals, a dedicated financial team is needed to manage billing, offer structured payment plans, and reduce potential debt from the high out-of-pocket burden.
Proactive Seasonal Resource Allocation	Use the identified seasonal peaks (**May** for overall admissions, **Summer/Fall** for Dengue) to adjust staffing, stock key supplies (e.g., Dengue test kits), and allocate beds proactively.



7. Next Steps for Advanced Analysis

1. **Cost Model Development:** Build a regression model to predict `Billing_Amount` based on `Medical_Condition`, `LOS_Days`, `Age`, and `Insurance_Standardized`.
2. **Clustering/Segmentation:** Use clustering algorithms (e.g., k-means) to segment patients based on treatment complexity (LOS and Cost) to identify distinct operational profiles (e.g., Inpatient vs. Outpatient vs. High-Acuity).
3. **Survival Analysis:** Apply survival analysis techniques (e.g., Kaplan-Meier) to the date data to model the probability of patient discharge over time, segmented by `Medical_Condition` and `Age`.