DATA_ENG 300_HW2_HealthcareDatabase
Marcus Van Mieghem
May 9, 2025

Part I

Analysis Question 1: Medication Usage Patterns by Ethnic Group

To create a summary of drug types and their total amount used by ethnicity, I determined drug type using the form_unit_disp field. This allowed me to classify prescriptions into categories such as Tablet/Capsule, Liquid, Injection/IV, Inhaled, Topical, Rectal, and Eye. I also attempted to do this by route but I found form_unit_disp_field to be a better alternative. In the uploaded code, both methods can be seen.

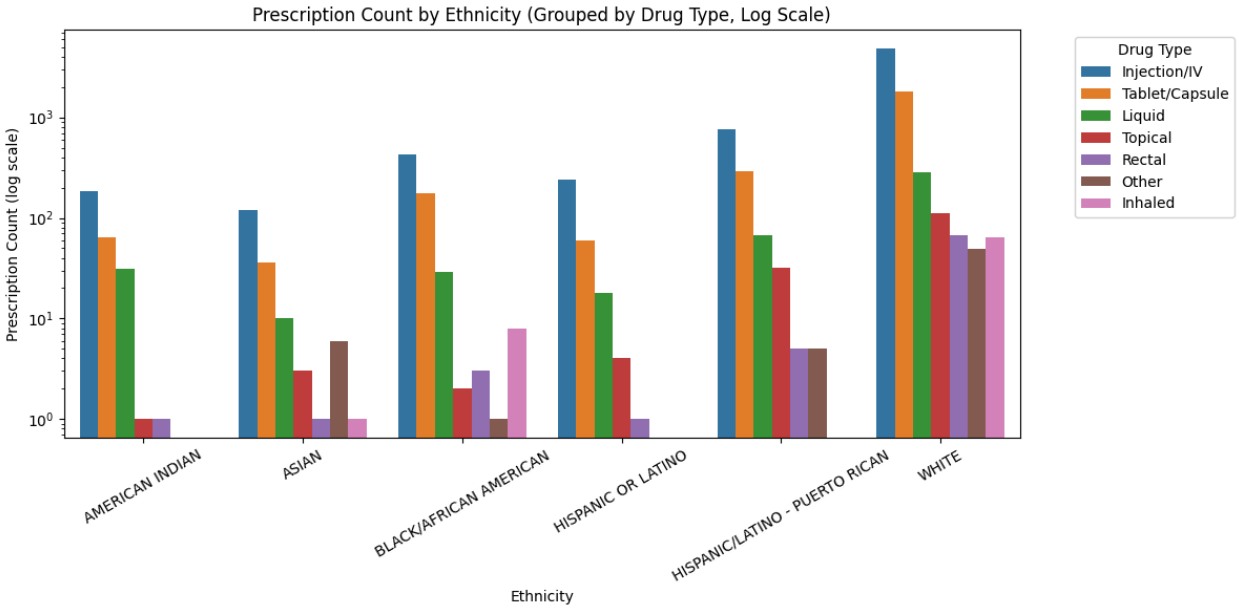
To quantify usage, I followed a recommendation from Professor Chan and used the number of occurrences as a proxy for total amount. While this approach does not capture actual dosage, it has the advantage of standardizing across patients, removing variability due to individual dosage requirements.

Some form_unit_disp entries, such as 'dose', were ambiguous and did not provide enough information to reliably classify the drug. These cases were grouped under Other. Additionally, one prescription had a missing form_unit_disp value and was excluded from the analysis, which accounts for the one-record discrepancy between the total number of prescriptions in the dataset and the total counted in the summary. There were also ethnicities such as "OTHER", "UNABLE TO OBTAIN", "UNKNOWN/NOT SPECIFIED", I did not include these in my analysis.

This query summarizes the number of prescriptions by drug type and patient ethnicity using the form_unit_disp field to define drug types. It begins by joining the PRESCRIPTIONS and ADMISSIONS tables on both subject_id and hadm_id to accurately match each prescription to the correct patient and their ethnicity. A CASE statement is then used to classify specific form_unit_disp values into broader drug type categories such as “Tablet/Capsule,” “Injection/IV,” “Inhaled,” and “Topical.” Only prescriptions with clearly defined form values are included in the analysis through a WHERE filter. The query then groups the data by ethnicity and drug type and counts the number of occurrences in each group. Finally, the results are ordered by ethnicity and descending prescription count to highlight the most commonly prescribed drug types for each group.

	ethnicity	drug_type	prescription_count
0	AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	Injection/IV	185
1	AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	Tablet/Capsule	64
2	AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	Liquid	31
3	AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	Topical	1
4	AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	Rectal	1

Visualizing the results:



To better visualize differences in prescription patterns across ethnicities, I used a log scale for the y-axis of the bar plot. This was necessary because the dataset is heavily skewed toward White patients, who make up the majority of the sample (74 of the 100 distinct patients with known ethnicity). Other ethnic groups have far fewer patients: Black/African American (6), Asian (2), Hispanic or Latino (2), Puerto Rican (1), and American Indian (1). For clarity, I excluded categories with unclear or ambiguous classification: "OTHER", "UNABLE TO OBTAIN", and "UNKNOWN/NOT SPECIFIED".

In the resulting visualization, we see that Injection/IV drugs are the most commonly prescribed type across every ethnicity, followed by Tablet/Capsule and then Liquid forms. For most ethnic groups, Topical drugs are the next most frequent, particularly among White, Hispanic, and Asian patients. Notably, White patients show a relatively balanced distribution between Topical, Rectal, and Inhaled drugs, a pattern which is not seen in the other groups. It is also worth noting that the dataset includes 10,397 prescriptions across just 100 patients, meaning there are, on average, around 100 prescriptions per patient. This shows the high frequency of medication administration in a hospital environment.

It's important to note that this analysis is based on an assumption: I classified drug types using the form_unit_disp field in the prescription data. While this field provides insight into how a drug is administered (e.g., tablets, vials, patches), it is not a perfect proxy for true drug type, and some ambiguity remains. For consistency and clarity, I excluded undefined or difficult-to-interpret units (e.g., 'dose', None) from the analysis. If this analysis were being used in high stake situations, I would deem it necessary to go one by one through the rows and make sure all data is correct.

Analysis Question 2: Procedures performed on patients by age group

To analyze procedures by age group, I first calculated each patient's age at the time of admission. This was done by computing the day difference between the patient's date of birth (dob) and their admission time (admittime), and dividing that result by 365.25 to account for leap years. The age was then floored to the nearest whole number:

Note: This calculation is an approximation and may not be accurate to the exact day (as it accounts for leap years), but it provides a reliable estimate for grouping. Once again, if this analysis were being used with high stakes, I would find a more accurate way to calculate age.

This query that I used then retrieves the short title for each procedure by matching the procedure code with its description in the reference table. It then calculates the patient's age at admission and classifies them into one of five age groups. For privacy reasons, patients aged over 90 have their age recorded as 300 in the dataset, although these are still grouped in the 80+ age group. The query groups results by both age group and procedure title, then orders them by age group and descending procedure frequency. The procedures are then ranked within each age_group. Then only the three highest ranked procedures are shown in the final output (after being converted to a dataframe). For example:

	procedure_title	age_group	procedure_count	procedure_rank
0	Venous cath NEC	0–19	3	1
1	Skin closure NEC	0–19	2	2
2	Packed cell transfusion	0–19	1	3

There were many ties for third rank of procedures for 0-19 year olds and 20-49 year olds. The other age groups did not have any ties. Here is a chart visualizing the results.

Age Group	Rank	Procedure Title	Count
0–19	1	Venous catheterization, not elsewhere classified	3
	2	Closure of skin and subcutaneous tissue of other sites	2
	3	Other diagnostic procedures on brain and cerebral meninges	1
		Note: 17 other procedures tied with 1 count each	
20–49	1	Venous catheterization, not elsewhere classified	8
	2	Enteral infusion of concentrated nutritional substances	7
	3	Percutaneous abdominal drainage	6
		Note: 2 other procedures also had 6 counts	
50–79	1	Venous catheterization, not elsewhere classified	26
	2	Enteral infusion of concentrated nutritional substances	22
	3	Transfusion of packed cells	13
80+	1	Venous catheterization, not elsewhere classified	19
	2	Transfusion of packed cells	13
	3	Insertion of endotracheal tube	8

Across all age groups, venous catheterization was the most common procedure, which makes sense as the need to deliver fluids, medications, or blood products directly into the bloodstream is very frequent.. In the 0–19 group, most procedures occurred only once, suggesting a wide range of less frequent interventions. For older age groups (20–49, 50–79, 80+), common procedures included nutritional support, transfusions, and airway management, highlighting an increase in supportive care with age. These results make sense as older patients usually need more intensive care.

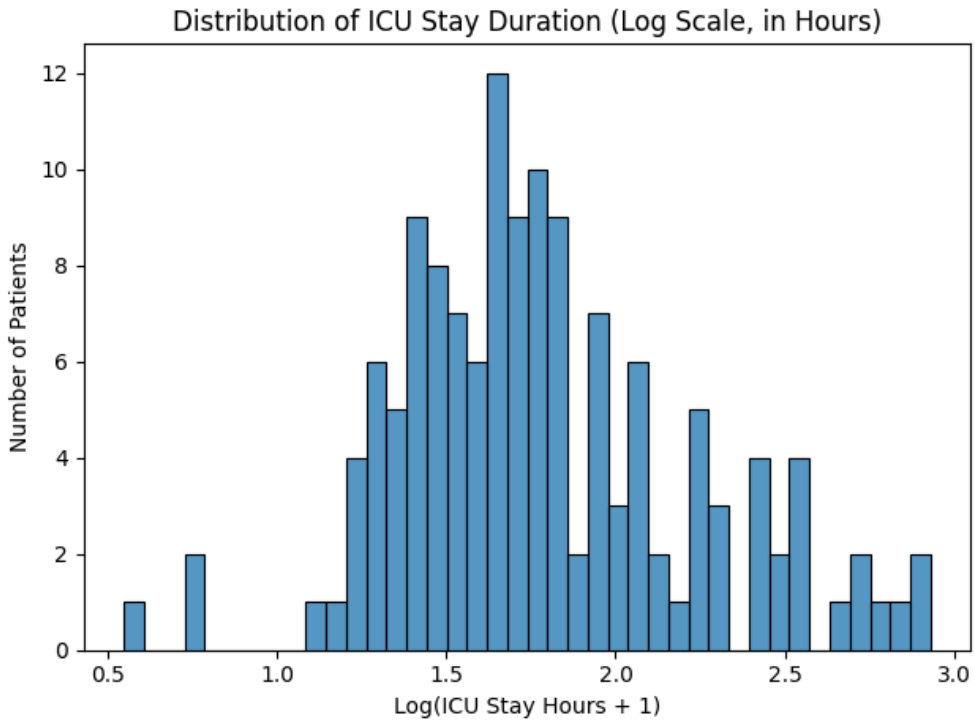
Analysis Question 3: ICU Duration Summary

Preface: After completing this question, I realized that the dataset includes a column for Length of Stay (LOS), which could have simplified the analysis. In hindsight, I would have used this column directly. However, I designed an alternative approach to calculate ICU stay duration, and it produced consistent results with what the LOS column would have yielded.

In order to calculate ICU duration I took the difference between the intime and outtime within ICUSTAYS. In order to calculate the difference, the query calculates the second difference and then divides by 3600 seconds to get hours. The final output (after being converted to a dataframe) shows subject_id, hadm_id, icustay_id, intime, our time, gender, ethnicity, icu_stay_hours and log_icu_stay_hours. For example first few rows of output:

	subject_id	hadm_id	icustay_id	intime	outtime	gender	ethnicity	icu_stay_hours	log_icu_stay_hours
0	10006	142345	206504	2164-10-23 21:10:15	2164-10-25 12:21:07	F	BLACK/AFRICAN AMERICAN	39.181111	1.593077
1	10011	105331	232110	2126-08-14 22:34:00	2126-08-28 18:59:00	F	UNKNOWN/NOT SPECIFIED	332.416667	2.521683
2	10013	165520	264446	2125-10-04 23:38:00	2125-10-07 15:13:52	F	UNKNOWN/NOT SPECIFIED	63.597778	1.803442
3	10017	199207	204881	2149-05-29 18:52:29	2149-05-31 22:19:17	F	WHITE	51.446667	1.711357
4	10019	177759	228977	2163-05-14 20:43:56	2163-05-16 03:47:04	M	WHITE	31.052222	1.492093

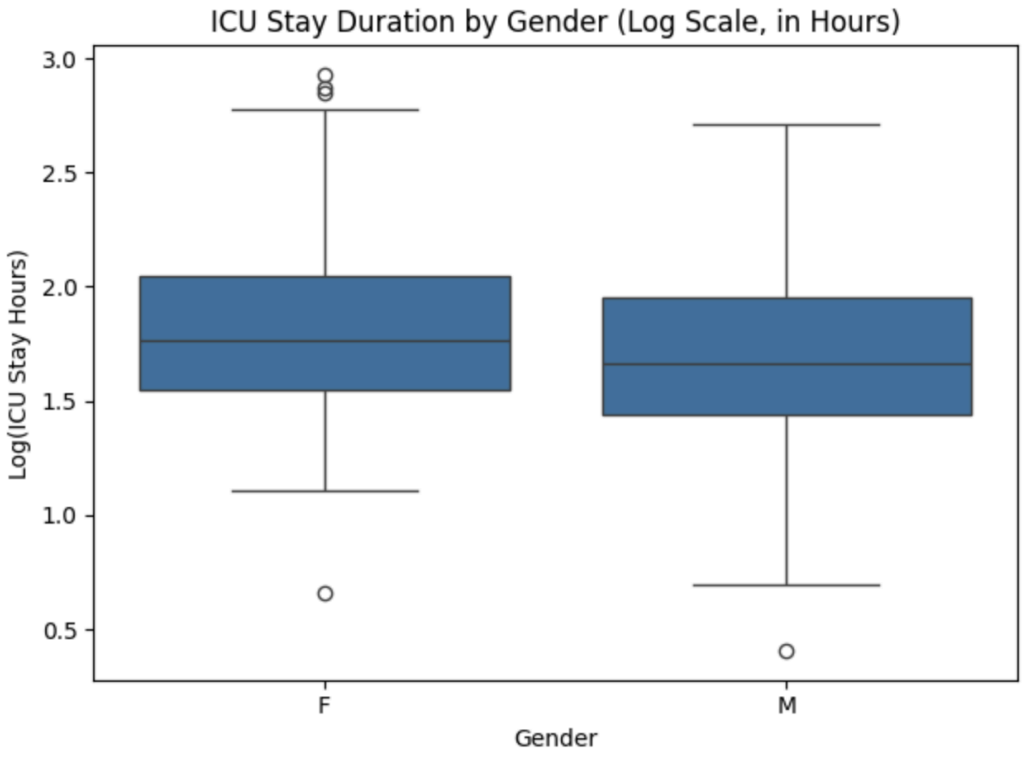
Visualizing the results of overall ICU stay duration:



The distribution of ICU stay durations, when log-transformed, appears fairly although definitely not perfectly normal. Log transformation was applied to better visualize the data, as the original distribution was highly skewed due to a small number of patients with extremely long ICU stays (those who may have had severe health issues). Without the log scale, the ICU stay ranged from 2.54 hours to 849.75 hours, with an average of 106.85 hours. The log transformation helps

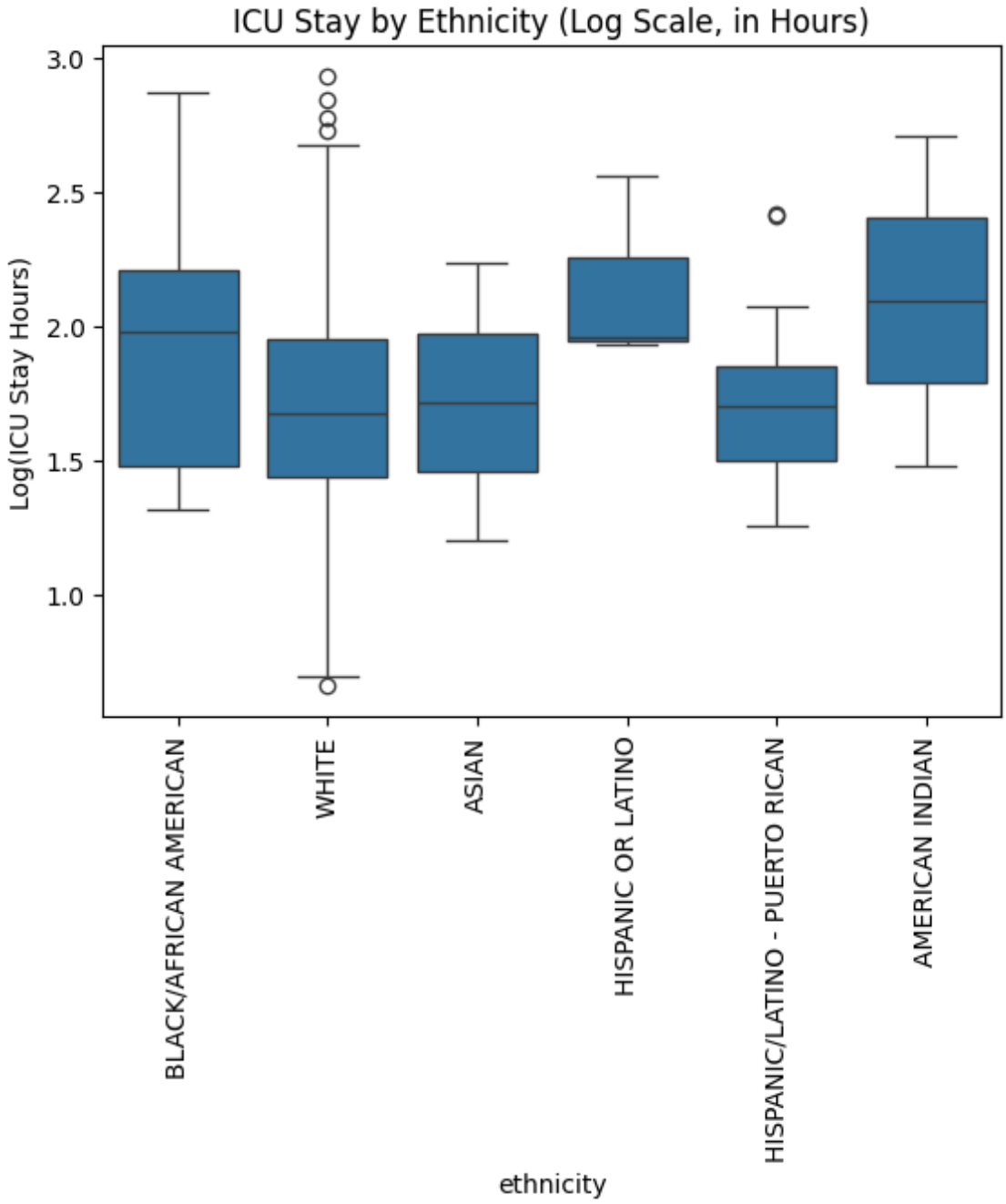
compress this range and reveal the underlying pattern of the majority of patient stays more clearly.

Now visualizing the results but this time separating by gender:



When comparing ICU stay durations by gender, females show a slightly higher median stay of 57.7 hours, compared to 46.2 hours for males. Females also had a longer maximum stay (948 vs. 513 hours). As expected, given that gender should not make much of a difference in ICU stay duration, the overall difference is small.

Now finally looking at ICU stay duration by ethnicity:



It appears that White individuals have the lowest median ICU stay, while American Indian patients have the highest median and the widest spread in ICU stay duration. There are also many outliers among White patients, which may be due to their much larger representation in the dataset (74 patients) compared to other groups. For example, there are only 2 Asian, 2

Hispanic, and 1 American Indian patients, making the distribution for those groups less reliable. This makes the chart somewhat surprising at first glance, since we wouldn't expect ICU stay to vary significantly by ethnicity given standardized care. However, the differences are likely driven more by small sample sizes in underrepresented groups than anything else. For clarity, I excluded ethnicities with unclear or ambiguous classification: "OTHER", "UNABLE TO OBTAIN", and "UNKNOWN/NOT SPECIFIED".

Part II

Acknowledgment
I confirm that no copies of the AWS credentials file are stored on any publicly accessible location, nor have I shared the file with anyone outside of DATA_ENG 300 (Spring 2025).
Signed: **Marcus Van Mieghem**

For Part II of the assignment, I focused on recreating the three analysis questions using denormalized tables in Cassandra (I created a keyspace ‘ueb9720_hw2‘), following the professor’s clarification that while no aggregations or joins are allowed in Cassandra, we are permitted to perform post-processing in Python and reuse intermediate joined data from Part I. For each question, I created a dedicated table in Cassandra containing raw, non-aggregated data structured to match the needs of the corresponding analysis. I then uploaded these datasets row-by-row from CSV files using session.execute() within a Jupyter notebook hosted on my EC2 instance. After confirming that all rows were successfully inserted, I ran simple SELECT * queries from each table and performed filtering, grouping, and visualization in Python using pandas and seaborn. The analysis and results exactly mirror those from Part I, including the bar plot of prescription drug type by ethnicity, the top three procedures by age group, and the ICU stay distributions by gender and ethnicity, and are fully documented in the notebook. As per the instructions and class guidance, all code and outputs for extraction and analysis are included in the Jupyter notebook submission and not repeated here.

Generative AI disclosure
I used Generative Artificial Intelligence (ChatGPT) to support parts of the coding for this assignment. Specifically, I used ChatGPT to:

- Create all plots in analysis question 1 2 and 3. Ai is very good at creating plots which can sometimes be very tedious.
    - Prompt: ‘I have exported the following columns to a dataframe x x x, can you help me make a plot with x on the horizontal axis and y on the vertical axis
- To insert data into the cassandra tables on part II of the homework. I was not sure of the best way to do this.
    - Prompt: ‘I have a dataframe with these columns, can you help me insert row by row into a cassandra table.’
- In analysis question 2, on part I, I was not sure how to rank the procedures in the query.
    - Prompt: ‘I have this query but I am not sure how to rank the procures by occurrence in age, can you help me’