Homework 2: Two ways with MIMIC-III

Written answers

Question 1- Create a summary of type of drugs and their total amount used by ethnicity. Report the top usage in each ethnicity group. *You may have to make certain assumptions in calculating their total amount.*

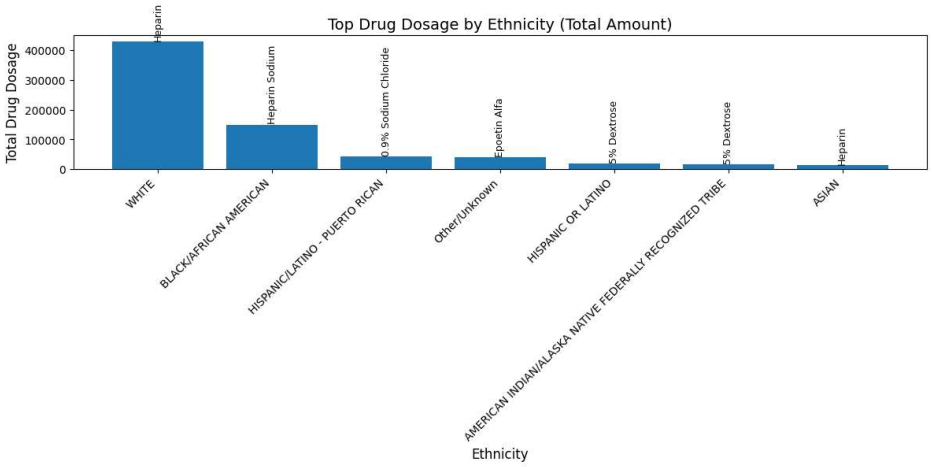
Part I:

(a) In the docker container

(b) I have written a series of smaller queries to answer question 1. To find the total drugs and their amount used by ethnicity, the query first groups together the `"UNKNOWN", "UNKNOWN/NOT SPECIFIED", "UNABLE TO OBTAIN"`, and `"OTHER/UNKNOWN"` categories into one category called `"Other/Unknown"` to get rid of redundant data. This cleaned data is stored as `admissions_cleaned`. The query then, joins prescriptions and admissions using `hadm_id` to map each drug to the patient's ethnicity and stores it as `prescriptions_cleaned`. Then, assuming that `dose_val_rx` is the amount of a drug used, the next query sums up this value while grouping by ethnicity and drug to calculate the total amount of each drug used by ethnicity and stores this as a new table. Finally, `ROW_NUMBER()` is used to get the top usage in each ethnicity group based on the previously calculated total dose value.

(c) In the docker container

(d) As can be seen in the graph below, Heparin is the most used drug among White people with a value of 427700 which is by far the highest number of doses in all the ethnicity groups. The next highest number of drug doses is 150000 for Heparin Sodium for people of Black or African American ethnicity. Puerto Rican Hispanic/Latino people use 0.9% Sodium Chloride the most. Hispanic/Latino people use 5% dextrose the most while Asians use Heparin the most. This suggests that Heparin and its variants seem to be the most commonly used drugs across ethnicities.



Part II:

(a) In the docker container

(b) In the docker container

(c) I did not use Cassandra for aggregation but used Pandas instead. I wanted to use functions such as limit and groupby that are not supported directly by Cassandra. I found it simpler to use pandas to summarize and aggregate data.

(d) The resulting data table proves that the extraction produces the expected results.

	ethnicity_group	drug	total_amount
5	AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	5% Dextrose	16900.0
3	ASIAN	Heparin	15000.0
1	BLACK/AFRICAN AMERICAN	Heparin Sodium	150000.0
6	HISPANIC OR LATINO	5% Dextrose	19950.0
4	HISPANIC/LATINO - PUERTO RICAN	0.9% Sodium Chloride	43663.0
0	Other/Unknown	Epoetin Alfa	40000.0
2	WHITE	Heparin	427700.0

Question 2 - Create a summary of procedures performed on patients by age groups (<=19, 20-49, 50-79, >80). Report the top three procedures, along with the name of the procedures, performed in each age group.

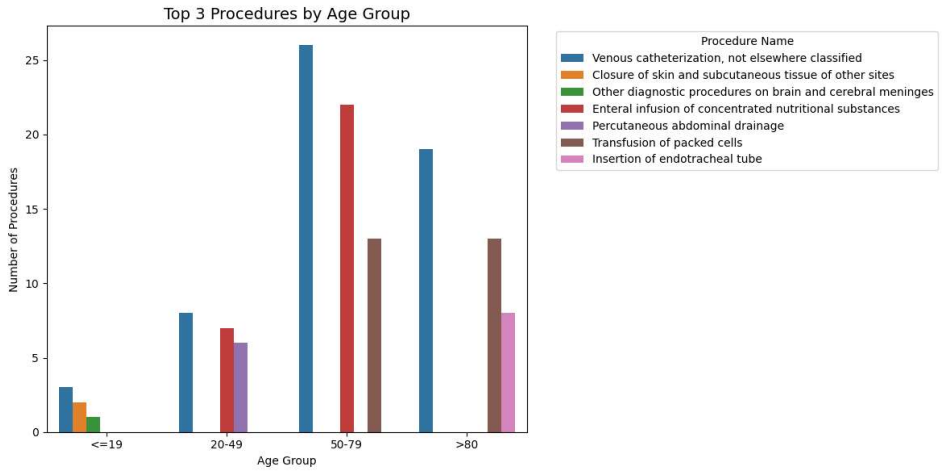
Part I:

(a) In the docker container

(b) This series of queries first calculates the age of the patient at the time of admission by calculating the difference between the admission time and the date of birth using the `DATE_PART` function. A new table called `patient_age` is created by joining the `admissions` and `patients` tables on `subject_id`. Another table called `patient_age` is created to divide the patients into the required four categories by age. These two tables are joined on the `hadm_id` and `subject_id` to create another table called `procedures_by_age` to connect the procedure to the patient for each distinct hospital visit. The table `procedure_counts` groups the `procedures_by_age` data using `icd9_code`. To improve readability and display the long titles of the procedures, a left join on `icd9_code` is performed. Finally, a row number is assigned to each procedure within age groups based on the highest number of `proc_count`. Order by is used to find the top three procedures for each `age_group` based on this calculated rank.

(c) In the docker container

(d) As can be seen in the graph below, Venous catheterization is the most performed procedure across all the age groups. For older age groups, above the age of 50, Transfusion of packed cells and Enteral infusion of concentrated nutritional substances appear consistently in the top three procedures. The number of procedures as well as the complexity of procedures performed increases significantly with age and is the highest in the 50-79 age group. This could also be due to the fact that there are lesser patients in the >80 category. For the age group below 19, the other two procedures with very low counts are Closure of skin and subcutaneous tissue of other and Repair of vertebral fracture indicating that there are very few patients below the age of 19 who underwent procedures and the ones who did had likely gotten into accidents.



Part II:

(a) In the docker container

(b) In the docker container

(c) I used pandas instead of Cassandra for aggregating the data so that I could directly access pandas specific functions to group the data and summarize it.

(d) The results were as expected and matched the duckDB results.

	age_group	icd9_code	procedure_name	proc_count	rn
0	<=19	3893	Venous catheterization, not elsewhere classified	3	1
1	<=19	8659	Closure of skin and subcutaneous tissue of oth...	2	2
2	<=19	0118	Other diagnostic procedures on brain and cereb...	1	3
3	20-49	3893	Venous catheterization, not elsewhere classified	8	1
4	20-49	966	Enteral infusion of concentrated nutritional s...	7	2
5	20-49	5491	Percutaneous abdominal drainage	6	3
6	50-79	3893	Venous catheterization, not elsewhere classified	26	1
7	50-79	966	Enteral infusion of concentrated nutritional s...	22	2
8	50-79	9904	Transfusion of packed cells	13	3
9	>80	3893	Venous catheterization, not elsewhere classified	19	1
10	>80	9904	Transfusion of packed cells	13	2
11	>80	9604	Insertion of endotracheal tube	8	3

Question 3 - How long do patients stay in the ICU? Is there a difference in the ICU length of stay among gender or ethnicity?

Part I:

(a) In the docker container

(b) This query first creates a table called `icu_stays_duration` to calculate the duration of the ICU stay for each patient, in hours, using the `DATE_DIFF` function. It uses this function to subtract the intime from the outtime to calculate the number of hours a patient stays in the ICU. To analyze if there is a difference in gender and ethnicity, the duration table is joined with `patients` and `admissions` tables on the `subject_id` and `hadm_id` respectively and cleaned by converting everything to lowercase and removing leading and trailing white spaces. The cleaned data is stored in `icu_cleaned_ethnicity`. To analyze differences in gender, the data in `icu_cleaned_ethnicity` is grouped by gender and the `AVG` and

`MEDIAN` hours are calculated by gender. The same thing is done for ethnicity. I decided to calculate both the mean and the median because the median is a better indicator if outliers are present. If some patients have really long stays, the mean can get skewed.

(c) In the docker container

(d) We can see from the results that females have a higher average ICU stay duration compared to males. Females spend around 132.98 hours on average with the median being 58 hours, while males spend around 84.32 hours on average with the median being 46 hours. The average and the median is higher for females.

gender varchar	num_stays int64	avg_hours double	median_hours double
F	63	132.98	58.0
M	73	84.32	46.0

As for the ethnicity, White patients had the highest number of ICU stays by far but had the lowest average ICU stay duration of 99 hours. Black patients had the highest average ICU stay duration of 184 hours with a suprisingly high median of 95 hours, but had fewer ICU stays compared to White patients and other ethnicities. Asians had the least number of ICU stays with an average length of 93 hours. The Hispanic patients had a similar average length of 94 hours but a median of 63.5 hours. A significant number of ICU stays were ambiguous ethnicity patients but they had an average ICU stay length of 132 hours with a median of 64 hours.

ethnicity_group varchar	num_stays int64	avg_hours double	median_hours double
Other/Unknown	17	131.24	64.0
Asian	2	93.0	93.0
Black	7	184.29	95.0
White	92	99.11	47.5
Hispanic	18	94.89	63.5

Part II:

(a) In the docker container

(b) In the docker container

(c) I used pandas for the post extraction analysis because I found it easier to group and summarize data using the pandas mean and median functions - that isn't possible directly in Cassandra.

(d) The results for this question are for the most part as expected. However, we see that there is a slight difference in the average hours spent as compared to the results from the duckDB analysis. These minor discrepancies could be due to rounding differences, precision for floating point numbers, and missing rows during uploading to Cassandra. Since these are very minor differences, they will not affect the overall interpretation of the results.

Signed Acknowledgement:

> *No copies of the AWS crendentials file is stored on any publicly accessible location, nor is the file in any way shared with anyone outside of DATA_ENG 300 (Spring 2025).*

Signature: Siddhika Swarup

Generative AI Disclosure

(1) I used GenAI for help with syntax for sql functions and pandas operations, and in depth help specifically for uploading data to Cassandra tables and clarifying concepts like ROW_NUMBER() and how PARTITION BY works.

(2) ChatGPT 4o.

(3) The prompts I used to get the results are as follows:
Can you give me sql functions to rank data?
How does PARTITION BY work and does it make sense to use it in this query?
Some rows of data are missing when I upload my data to the Cassandra table. Can you help me fix this?