DE300 HW2 Analysis Answers - Nicky Williams

Part I (Relational)

1. Answer the following *analysis questions* along with your queries.
   For each question, provide the following four pieces of information:
   A. the SQL query,
   B. a brief explanation of the query (i.e., what operations are performed by the major parts of the query),
   C. the first several lines of your resulting table, and
   D. a summary of your findings. If it benefits to use a graph, include your graph at the end of your answer, with clear labels and captions.

The analysis questions are:

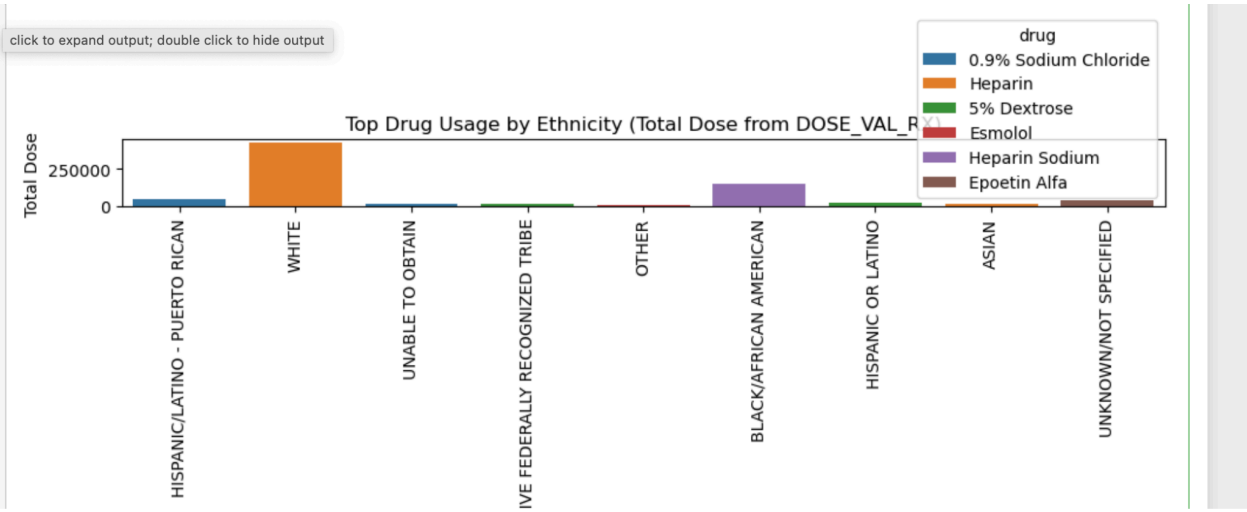
1. Create a summary of types of drugs and their total amount used by ethnicity. Report the top usage in each ethnicity group. *You may have to make certain assumptions in calculating their total amount.*
2. Create a summary of procedures performed on patients by age groups (<=19, 20-49, 50-79, >80). Report the top three procedures, along with the name of the procedures, performed in each age group.
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 Answers

1. Drug types and amount by ethnicity
   a. Query provided in code
   b. Query joins prescriptions with admissions on hospital ID to relate drug use with patient ethnicity. It filters through only the valid dose values using regex, casts them into numbers, sums the total per drug and ethnicity, and orders them in descending order to select the top drug by dosage from each ethnicity group.
   c. First several lines of resulting table:

	ethnicity	drug	total_dose
0	HISPANIC/LATINO - PUERTO RICAN	0.9% Sodium Chloride	43663.0
1	WHITE	Heparin	427700.0
2	UNABLE TO OBTAIN	0.9% Sodium Chloride	14800.0
3	AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	5% Dextrose	16900.0
4	OTHER	Esmolol	5000.0

d. Summary: 0.9% sodium chloride was the most used drug overall with both Puerto rican and American Indian ethnic groups having that as their most used drug. From the graph below we can also see that Heparin is greatly used by the white ethnic group and widely surpasses other top drugs in other ethnicities. Many ethnicities have relatively low top drug usage compared to White, African American, and Puerto Rican groups.



2. Procedures performed on patients by age group
    a. Query provided in code
    b. Query joins patients and admissions table on subject id, joins to procedures_id table on hospital id, and then joins d_icd_procedures on icd9_code. It translates birth date and admit time into age and creates 3 different groups for the age categories. I then identified the top 3 procedures per age group. Age groups were calculated by subtracting the date of birth from the admission date and then dividing by 365.

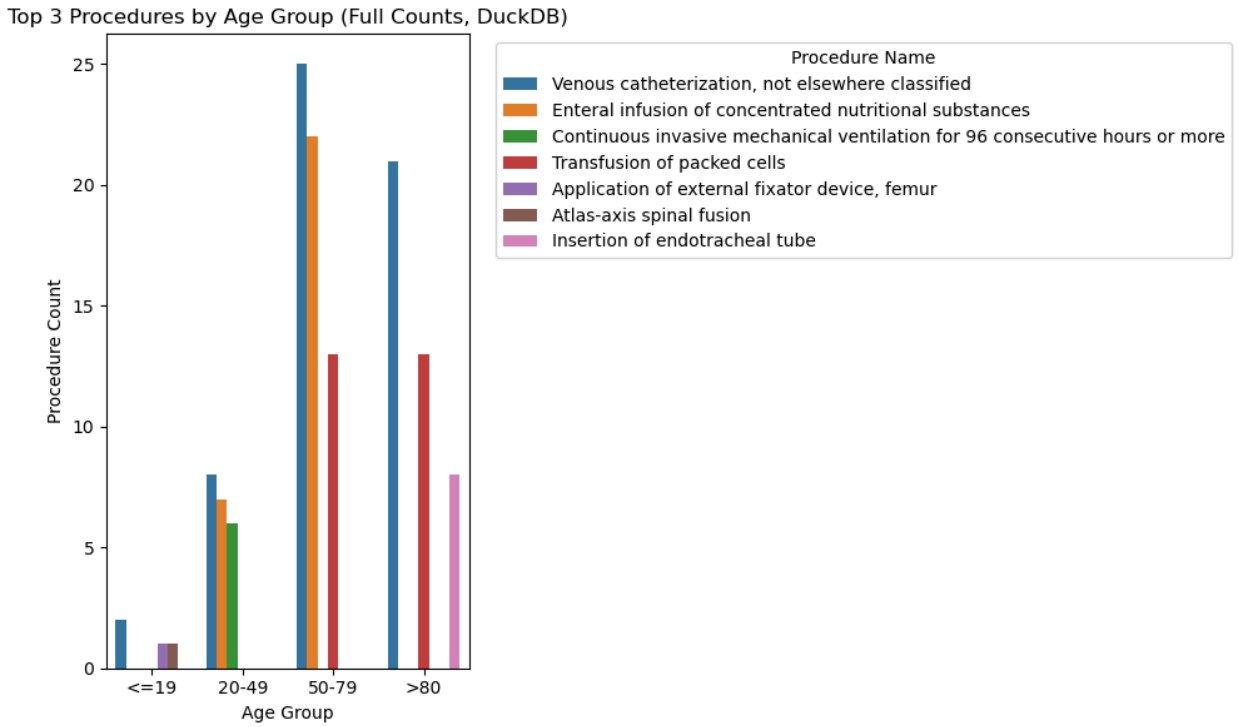
c. First several lines of resulting table:

1]:

	age_group	procedure_name	procedure_count
43	20-49	Venous catheterization, not elsewhere classified	8
12	20-49	Enteral infusion of concentrated nutritional s...	7
9	20-49	Continuous invasive mechanical ventilation for...	6
139	50-79	Venous catheterization, not elsewhere classified	25
72	50-79	Enteral infusion of concentrated nutritional s...	22
136	50-79	Transfusion of packed cells	13
157	<=19	Venous catheterization, not elsewhere classified	2
140	<=19	Application of external fixator device, femur	1
141	<=19	Atlas-axis spinal fusion	1
225	>80	Venous catheterization, not elsewhere classified	21
222	>80	Transfusion of packed cells	13
189	>80	Insertion of endotracheal tube	8

d. Summary: As seen in the graph below, procedures overall vary a lot between different age groups. Younger groups tended to have less procedures, with the count increasing over time and peaking in the 50-79 age bucket. However, across all age groups, venous catheterization was the most frequent procedure. Additionally, older age groups tended to have more invasive and supportive procedures while younger patients tended to have more specialized orthopedic

procedures.



3. ICU Stays
    a. Query provided in code
    b. Query computes ICU stay length in days (based on timestamp difference between admit and discharge date - divided by secs in a day) by joining admissions with icustays on hospital id and joining admissions to patients on subject id. It then groups by gender and ethnicity and orders descending.
    c. First several lines of table:

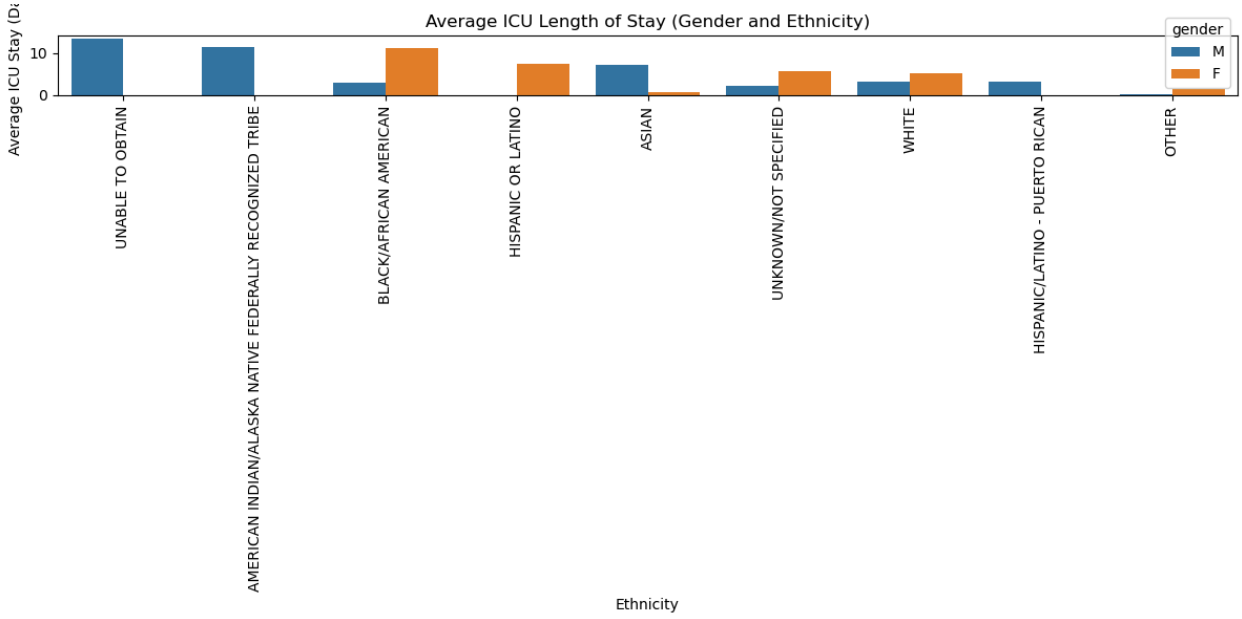
Ad

Ethnicity

	gender	ethnicity	avg_icu_stay_days
0	M	UNABLE TO OBTAIN	13.36
1	M	AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	11.34
2	F	BLACK/AFRICAN AMERICAN	11.20
3	F	HISPANIC OR LATINO	7.46
4	M	ASIAN	7.12

    d. Summary: Males generally had longer ICU stays in the highest-ranking categories. Patients with “unable to obtain” ethnicity had the longest ICU durations, which could mean that there were potential data quality issues or care complexity. African American and Hispanic females also had longer-than-average ICU stays. Overall

there are clear differences in length of ICU stays between both gender and ethnicity as can be seen in the image below:



Part II (Non-relational)

1. For each of the *analysis questions* above, provide the following four pieces of information:
    A. Design a Cassandra table for the specific analysis. Report your table creation query.
    B. Upload the data into the table to facilitate answering the question. Report your code for uploading the data.
    C. Report the query for extracting relevant data to answer the question. *You may choose to not aggregate within Cassandra. If so, indicate your post-extraction analyses and include the code for reaching the final answer.*
    D. Verify that the extraction produces the desired data

- Table creating query, code for uploading data, and query for extracting data for all 3 analysis questions provided in the code.
- Verified all extractions produce the desired data