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.*
    A. the SQL queries:

```
ethnicity_drug = conn.sql(
    """
        SELECT
            ethnicity,
            drug_type,
            COUNT(*) AS total_amount
        FROM PRESCRIPTIONS
        LEFT JOIN ADMISSIONS
        ON PRESCRIPTIONS.hadm_id = ADMISSIONS.hadm_id
        GROUP BY ethnicity, drug_type
        ORDER BY ethnicity, drug_type;
    """
).df()
```

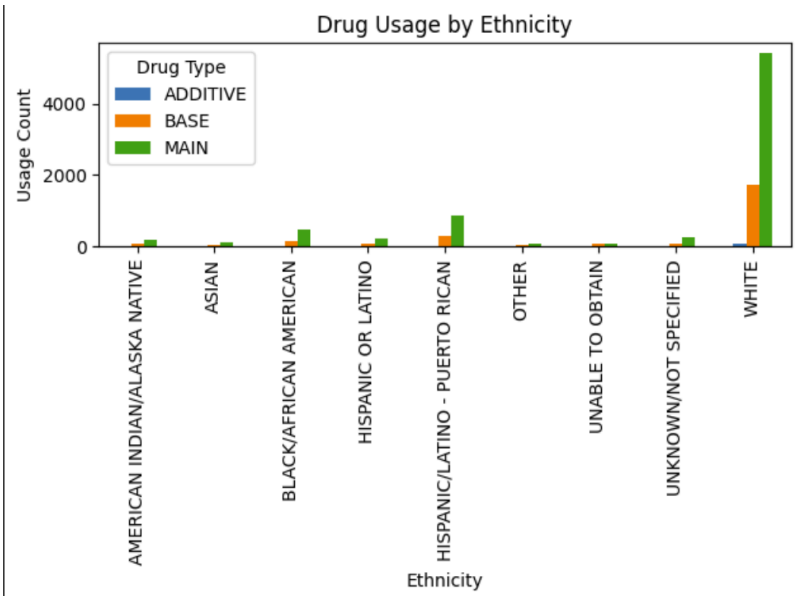
```
conn.sql(
    '''
    WITH drug_totals AS (
        SELECT
            ethnicity,
            drug_type,
            COUNT(*) AS total_amount
        FROM PRESCRIPTIONS
        JOIN ADMISSIONS
        ON PRESCRIPTIONS.hadm_id = ADMISSIONS.hadm_id
        GROUP BY ethnicity, drug_type
    ),
    max_totals AS (
        SELECT
            ethnicity,
            MAX(total_amount) AS max_total
        FROM drug_totals
        GROUP BY ethnicity
    )
    SELECT
        drug_totals.ethnicity,
        drug_totals.drug_type,
        drug_totals.total_amount AS frequency_of_usage
    FROM drug_totals JOIN max_totals ON drug_totals.ethnicity = max_totals.ethnicity AND drug_totals.total_amount = max_totals.max_total
    ORDER BY drug_totals.ethnicity;
'''
).df()
```

    B. a brief explanation of the query (i.e., what operations are performed by the major parts of the query):
        a. This analysis assumes that the total amount is the number of occurrences. In the first query, admissions and prescriptions are joined to have both ethnicity and drug_type in a single table. The number of rows is tracked and then is grouped by ethnicity and drug type so that the number of times a prescription has been obtained for different drug types for patients of different ethnicities is displayed.
        b. The second query builds off of the first query and now tracks the maximum total amount used by ethnicity such that the most highly used drug type is displayed in the select statement for each ethnicity.
    C. the first several lines of your resulting table:

D.

	ethnicity	drug_type	frequency_of_usage
0	AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	MAIN	200
1	ASIAN	MAIN	121
2	BLACK/AFRICAN AMERICAN	MAIN	476
3	HISPANIC OR LATINO	MAIN	226
4	HISPANIC/LATINO - PUERTO RICAN	MAIN	860
5	OTHER	MAIN	72
6	UNABLE TO OBTAIN	MAIN	89
7	UNKNOWN/NOT SPECIFIED	MAIN	245
8	WHITE	MAIN	5420

E. 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 caption:
   a. The bar graph shows that white patients have the highest total frequency of usage across all drugs and that all ethnicities use drugs of type main the most.



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.
   A. Please find my SQL queries in my Jupyter Notebook
   B. The first query creates a new column in the Admissions table representing the age group in which the patient belongs. It creates a case statement that assigns an age group based on the patient's age at admission (admission time minus birthdate) . In the next query, the age group, procedure title, and count of procedures are extracted from joining of procs_icd, d_icdprocs, and admissions. These counts are grouped first by age group and then by the procedure type. In order to extract the top three most common procedures from each age group, a row number is assigned to each

procedure within its age group. These are then sorted and the qualify statement ensures that only the top three procedures are outputted. To keep it easy to read, the data in the table is first ordered by age group, then procedure count, then procedure name.

C. See code for first few lines of table.
D. The bar graph generated demonstrates that across all three age groups, the venous cath NEC procedure is the most common. Additionally, it appears that patients under the age of 19 have the least number of recurring procedures, while patients between the ages of 50 and 79 have the most.

3. How long do patients stay in the ICU? Is there a difference in the ICU length of stay among gender or ethnicity?
    A. Please see the submitted Jupyter Notebook for specific queries.
    B. The first query adds a new column to icustays calculating the total length of stay for each patient in hours by subtracting their in time from out time. The second query selects the gender, ethnicity, and hours of stay from the joining of tables icustays, patients, and admissions. This combines all of the data needed for analysis into one table. The next query calculates the average number of hours of stay in the ICU by gender and ethnicity by aggregating the hours with the average function, joining the same tables listed above, and grouping by both gender and ethnicities so that the averages for ethnicities divided into two gender groups will be outputted. The next two queries perform the same actions with the only difference being that one groups by just gender and one groups by just ethnicity. In this way, the average times for the overall groups (gender and ethnicity) were outputted.
    C. Please see code for the first few lines of produced tables.
    D. The graphs produced give the following information. White patients have the widest range of hours spent in the ICU while American Indian and Unable to Obtain have the highest average number of hours spent in the ICU. Females have both a higher average and wider range of hours spent in the ICU compared to males. Females of most ethnicities tend to have longer ICU Stays than men except for Asian patients where males have much longer average stays. Black females have proportionally longer ICU stays than black males compared to other ethnicities. Of the explicitly specified ethnicities, American Indian patients have the longest average stays and Puerto Rican patients have the shortest average stay in the ICU.

For Part II, please see the attached .ipynb for my responses.

2. Please sign and acknowledge the following statement:

> *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).*

Rachel Silverman

Generative AI statement: ChatGPT was used mainly in resolving errors, specifically in setting up my keyspace on my EC2 instance and finding small bugs that occurred in my produced code. Therefore, most of my prompts were simply showing my error messages to ChatGPT and explaining the situation to ask for guidance. ChatGPT also helped me determine how to partition by age group in the first question. Specifically, this line of code, QUALIFY ROW_NUMBER() OVER (PARTITION BY age_group ORDER BY procedure_count DESC) <= 3, was produced with the help of Generative AI.