

## PART 1

- 1) For analysis question 1, I chose this query to group and count the amount of times a drug is prescribed to a certain ethnicity. In order to access the ethnicity and drug information, the subject ID and hospital admission ID are used to eliminate repeat entries and ensure that the ethnicity matched the person who matched the prescription with the drug. Below is the query, the table received from it, and the conclusion I reached. The last image shows that couples of ethnicities seem to share common prescriptions but the black/african american patients seem to get insulin more frequently and whites get potassium chloride.

```
2) """
3)     SELECT ADMISSIONS.ethnicity,
4)           PRESCRIPTIONS.drug,
5)           count(*)
6)     FROM PRESCRIPTIONS
7)     JOIN ADMISSIONS
8)           ON PRESCRIPTIONS.subject_id = ADMISSIONS.subject_id
9)           AND PRESCRIPTIONS.hadm_id = ADMISSIONS.hadm_id
10)    GROUP BY ADMISSIONS.ethnicity, PRESCRIPTIONS.drug
11)    """
```

ethnicity varchar	drug varchar	count_star() int64
WHITE	Senna	66
WHITE	Aspirin	44
WHITE	Tamsulosin	5
WHITE	Docusate Sodium	67
WHITE	Sodium Chloride 0.9% Flush	151
WHITE	Albuterol Inhaler	14

```
# AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN... ==== 5% Dextrose
# ASIAN ==== DSW
# BLACK?AFRICAN AMERICAN ==== Insulin
# HISPANIC OR LATINO ==== 5% Dextrose
# HISPANIC/LATINO - PUERTO RICAN ==== 0.9% Sodium Chloride
# OTHER ==== NS
# UNABLE TO OBTAIN ==== 0.9% Sodium Chloride
# UNKNOWN/NOT SPECIFIED ==== DSW
# WHITE ==== Potassium Chloride
```

2) Below is an example of the search queries I did for question 2. IN order to make it possible, I added a column “age” to the ADMISSIONS data. This allowed me to query based on age so each query only differed by the age condition for the JOIN. The results showed that venous cath NEC is extremely common at all ages but procedures are far more common in the 50-79 age range. IN addition, the packed cell transfusion seems to be common among the 50-79 and 80+ age groups.

```

"""
SELECT D_ICD_PROCEDURES.short_title,
       count(*)
FROM D_ICD_PROCEDURES
JOIN PROCEDURES_ICD
  ON PROCEDURES_ICD.icd9_code = D_ICD_PROCEDURES.icd9_code
JOIN ADMISSIONS
  ON PROCEDURES_ICD.subject_id = ADMISSIONS.subject_id
  AND PROCEDURES_ICD.hadm_id = ADMISSIONS.hadm_id
  AND ADMISSIONS.age <= 19
GROUP BY D_ICD_PROCEDURES.short_title
"""

```

	short_title	count_star()		short_title	count_star()
4	Venous cath NEC	2	16	Venous cath NEC	9
1	Cl fx reduc-femur	1	23	Enteral infus nutrit sub	7
0	Incision of lung	1	4	Percu abdominal drainage	6

  

	short_title	count_star()		short_title	count_star()
27	Venous cath NEC	25	20	Venous cath NEC	20
46	Enteral infus nutrit sub	22	10	Packed cell transfusion	13
18	Packed cell transfusion	13	49	Insert endotracheal tube	8

3) For question 3, I needed to add a column in ICU STAYS that held the days between the check in and check out of the patient. I was then able to join the icu stays with patients to examine the relation to gender and admissions for the relation to ethnicities. The query only changed in the selection, join, and grouping based on whether I wanted the ethnic divisions or the gender divisions. From the data I saw that women typically find themselves in the ICU longer and it can vary widely for different ethnicities. For ethnicities, it seems that whits and asians spend less days in the ICU on average when compared to african americans or hispanics.

```

"""
SELECT PATIENTS.gender,
       AVG(ICUSTAYS.days) AS average_stay_in_days
FROM ICUSTAYS

```

```

JOIN PATIENTS
    ON ICUSTAYS.subject_id = PATIENTS.subject_id
GROUP BY PATIENTS.gender
""")

```

gender varchar	average_stay_in_days double
F	5.476190476190476
M	3.5205479452054793

ethnicity varchar	average_stay_in_days double
UNKNOWN/NOT SPECIFIED	4.461538461538462
ASIAN	4.0
BLACK/AFRICAN AMERICAN	6.888888888888889
OTHER	1.0
HISPANIC OR LATINO	7.333333333333333
UNABLE TO OBTAIN	14.0
AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGNIZED TRIBE	11.5
WHITE	4.024590163934426
HISPANIC/LATINO - PUERTO RICAN	3.2666666666666666

## PART 2

- Overall had a very similar approach for part 2 but I used pandas dataframes to deal with any necessary merges and column creations. I then used insertions to fill in the cassandra tables with the proper data. While I have all of my code in hw2.ipynb, the queries/code can be found in hw2\_2.ipynb. I was able to get the same results for analysis question 1 but I got slightly different results in the other two questions.

## GENERATIVE AI DISCLOSURE

Used chatgpt for debugging and for help formatting print statements. I ran into quite a few errors with part 2 question 2 because of integer overflow caused by the fake dates used in the dataset. In order to get past that I used a code generated by chatgpt that solved the error.