Data Engineering 300

Homework 2 Write Up

Sabian Atmadja (sna3362)

Analysis Questions

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.***

The SQL query is found below:

```
SELECT
    ETHNICITY AS ETHNICITY,
    DRUG AS DRUG,
    AMOUNT
FROM (
    SELECT
        ADMISSIONS.ETHNICITY,
        PRESCRIPTIONS.DRUG,
        COUNT(*) AS AMOUNT
    FROM PRESCRIPTIONS
    JOIN ADMISSIONS ON PRESCRIPTIONS.SUBJECT_ID = ADMISSIONS.SUBJECT_ID
    WHERE DOSE_VAL_RX ~ '^[0-9]+(\\.[0-9]+)?$'
    GROUP BY ADMISSIONS.ETHNICITY, PRESCRIPTIONS.DRUG
) AS counts
QUALIFY AMOUNT = MAX(AMOUNT) OVER (PARTITION BY ETHNICITY)
ORDER BY ETHNICITY;
```

Figure 1: SQL Query for First Problem

The query basically does 2 things: it creates a subtable which is a table showing the amount of times any drug is prescribed to a particular ethnic group. I have assumed the total amount to be the number of times a drug is prescribed. From this subtable, you can query the maximum amount a drug is consumed per ethnicity. The qualify clause filters for the row(s) with the maximum prescription count per ethnicity. I have assumed that subject ID to the key to link the tables. The results can be found below:

ETHNICITY	DRUG	AMOUNT
AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	5% Dextrose	54
ASIAN	D5W	27
BLACK/AFRICAN AMERICAN	Insulin	60
HISPANIC OR LATINO	5% Dextrose	28
HISPANIC/LATINO - PUERTO RICAN	0.9% Sodium Chloride	1290
OTHER	NS	11
UNABLE TO OBTAIN	0.9% Sodium Chloride	28
UNKNOWN/NOT SPECIFIED	D5W	41
WHITE	Potassium Chloride	498

Figure 2: Results table for ethnicities and most often prescribed drugs

0.9% Sodium Chloride and 5% Dextrose are the most commonly prescribed drugs across different ethnic groups. One group that stands out is the Hispanic/Latino – Puerto Rican population, where 0.9% Sodium Chloride was prescribed 1,290 times — much higher than any other group. White patients most often received Potassium Chloride, with 498 prescriptions, which is still far behind the top count. Both Asian and Unknown/Not Specified groups had D5W as their most common drug, which might point to similar treatment practices. Meanwhile, groups like “Other” and “Unable to Obtain” had much lower prescription counts overall, which could mean they were underrepresented in the dataset.

Here we can see a clear discrepancy between different ethnic groups and the amount of times a drug is prescribed.

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.

```
SELECT D_ICDPROCS.LONG_TITLE AS PROCEDURE_NAME, COUNT(*) AS COUNT
FROM PROCS_ICD
JOIN ADMISSIONS ON PROCS_ICD.HADM_ID = ADMISSIONS.HADM_ID
JOIN PATIENTS ON PROCS_ICD.SUBJECT_ID = PATIENTS.SUBJECT_ID
JOIN D_ICDPROCS ON PROCS_ICD.ICD9_CODE = D_ICDPROCS.ICD9_CODE
WHERE FLOOR(DATE_DIFF('year', CAST(PATIENTS.DOB AS TIMESTAMP), CAST(ADMISSIONS.ADMITTIME AS TIMESTAMP))) <= 19
GROUP BY PROCEDURE_NAME
ORDER BY COUNT DESC
LIMIT 3;
```

Figure 3: SQL Query for second problem for most common procedures patients under 19

The following query was used, with the exception of the seventh line where the age limits were adjusted depending on the age group being analyzed. The query joins four tables: the procedures table (PROCS_ICD), admissions data (ADMISSIONS), patient demographics (PATIENTS), and a reference table for procedure names (D_ICDPROCS). These joins allow us to match procedure codes with their descriptions and associate each procedure with the patient's age at the time of admission. The patient's age is calculated using the difference in years between their date of birth and admission time. The query then groups the data by

procedure name and counts how many times each procedure occurred within a specific age group. Finally, it returns the top three most commonly performed procedures for that age group, sorted by frequency.

PROCEDURE_NAME	COUNT
Venous catheterization, not elsewhere classified	2
Incision of lung	1
Repair of vertebral fracture	1

Figure 4: Most common procedures for patients under the age of 19

PROCEDURE_NAME	COUNT
Venous catheterization, not elsewhere classified	9
Enteral infusion of concentrated nutritional s...	7
Percutaneous abdominal drainage	6

Figure 5: Most common procedures for patients between 20 and 49

PROCEDURE_NAME	COUNT
Venous catheterization, not elsewhere classified	25
Enteral infusion of concentrated nutritional s...	22
Transfusion of packed cells	13

Figure 6: Most common procedures for patients between 50 and 79

PROCEDURE_NAME	COUNT
Venous catheterization, not elsewhere classified	20
Transfusion of packed cells	13
Insertion of endotracheal tube	8

Figure 7: Most common procedures for patients over age of 80

Venous catheterization was the most common procedure across all age groups, suggesting it is a widely used intervention in inpatient care regardless of age. Transfusion of packed cells and insertion of endotracheal tubes were also frequent, especially in patients over 50, reflecting more intensive care needs in older populations. The youngest group (≤19) had significantly fewer procedures overall, possibly due to smaller sample size or less frequent hospitalization.

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

```
SELECT
    ROUND(AVG(
        DATE_DIFF('minute',
            CAST(ICUSTAYS.INTIME AS TIMESTAMP),
            CAST(ICUSTAYS.OUTTIME AS TIMESTAMP)
        ) / 60.0
    ), 2) AS AVG_ICU_HOURS
FROM ICUSTAYS
WHERE INTIME IS NOT NULL AND OUTTIME IS NOT NULL;
```

Figure 8: SQL query for average ICU stay for all patients

```
SELECT
    PATIENTS.GENDER,
    ROUND(AVG(
        DATE_DIFF('minute',
            CAST(ICUSTAYS.INTIME AS TIMESTAMP),
            CAST(ICUSTAYS.OUTTIME AS TIMESTAMP)
        ) / 60.0
    ), 0) AS AVG_ICU_HOURS
FROM ICUSTAYS
JOIN PATIENTS ON ICUSTAYS.SUBJECT_ID = PATIENTS.SUBJECT_ID
WHERE INTIME IS NOT NULL AND OUTTIME IS NOT NULL
GROUP BY PATIENTS.GENDER;
```

Figure 9: SQL query for average ICU stay by gender

```sql
SELECT
    ADMISSIONS.ETHNICITY,
    ROUND(AVG(
        DATE_DIFF('minute',
            CAST(ICUSTAYS.INTIME AS TIMESTAMP),
            CAST(ICUSTAYS.OUTTIME AS TIMESTAMP)
        ) / 60.0
    ), 0) AS AVG_ICU_HOURS
FROM ICUSTAYS
JOIN ADMISSIONS ON ICUSTAYS.HADM_ID = ADMISSIONS.HADM_ID
WHERE INTIME IS NOT NULL AND OUTTIME IS NOT NULL
GROUP BY ADMISSIONS.ETHNICITY
ORDER BY AVG_ICU_HOURS DESC;
```

Figure 10: SQL query for average ICU stay by ethnicity

The first query calculates the overall average ICU stay by taking the difference in minutes between ICU admission (INTIME) and discharge (OUTTIME), converting it to hours, and averaging across all patients. The second query joins the ICUSTAYS table with the PATIENTS table to compare average ICU duration by gender. The third query joins ICUSTAYS with ADMISSIONS to group by patient ethnicity and compute the average length of stay for each group. In all cases, we exclude ICU stays with missing time values to ensure accurate calculations.

AVG_ICU_HOURS
106.86

Figure 11: Average ICU stay for all patients

gender	AVG_ICU_HOURS
F	133.0
M	84.0

Figure 12: Average ICU stay by gender

ethnicity	AVG_ICU_HOURS
UNABLE TO OBTAIN	321.0
AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	272.0
BLACK/AFRICAN AMERICAN	184.0
HISPANIC OR LATINO	179.0
UNKNOWN/NOT SPECIFIED	118.0
WHITE	99.0
ASIAN	93.0
HISPANIC/LATINO - PUERTO RICAN	78.0
OTHER	22.0

Figure 13: Average ICU stay by ethnicity

The results show that the average ICU stay across all patients is 106.86 hours, or roughly 4.5 days. However, there are notable differences when broken down by gender and ethnicity. Female patients had a significantly higher average ICU stay (133 hours) compared to male patients (84 hours). Ethnic differences were even more pronounced: patients whose ethnicity was listed as “Unable to Obtain” had the longest average stay (321 hours), followed by American Indian/Alaska Native patients (272 hours) and Black/African American patients (184 hours). On the shorter end, Hispanic/Latino – Puerto Rican and Other groups had average stays of 78 and 22 hours, respectively. These differences may reflect variations in care needs, disease severity, or access to resources across demographic groups.

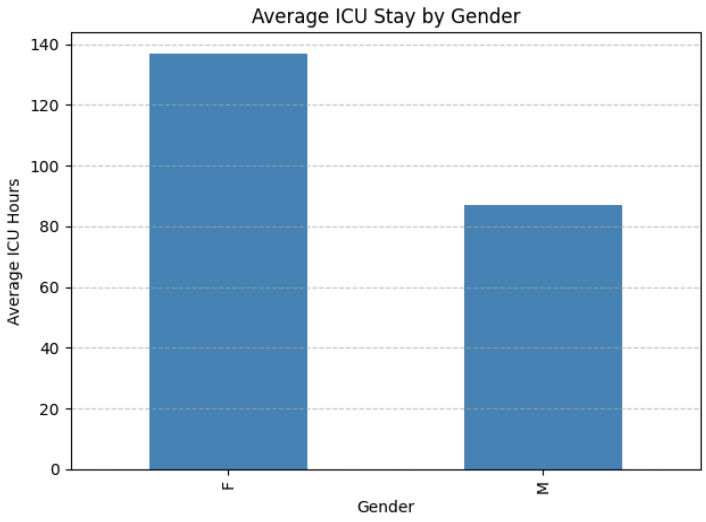

Figure 14: Average ICU stay by gender graph

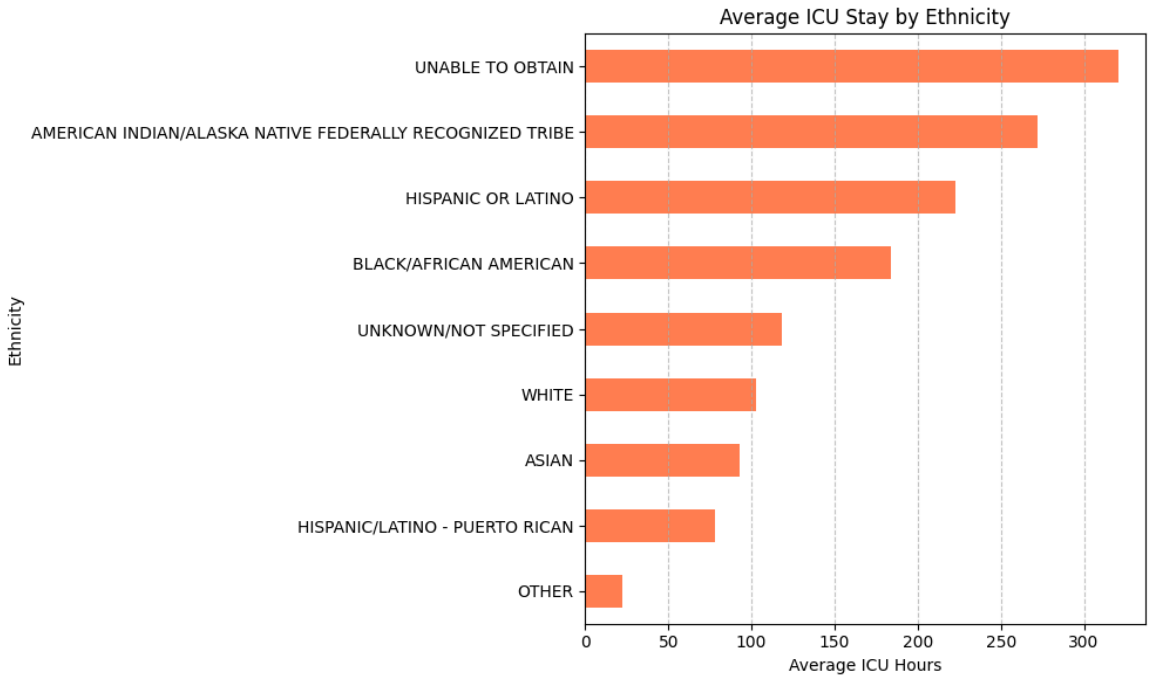

Figure 15: Average ICU stay by ethnicity graph

Cassandra Clarification

For the life of me, whenever I try to use pandas merge for any non-count like data, it will result in a different output. I have tried to remove duplicates and have asked Chat GPT. There is an inherent difference between how pandas merges (because it does not normalize) and SQL (which normalizes). I cannot seem to fix the issue but the results are similar. Please take a look and try for yourself.

Gen AI Disclosure

I used Chat GPT for a few things, namely on how to extract out the drug names for the first problem, debugging cassandra and why pandas' merge and SQL merge are different. I used several queries, namely:

1. I have this SQL query which effectively finds the amount of all drugs used by each ethnicity, how do I extract out the highest prescribed drug ethnicity from here?
2. Cassandra is not letting me upload the data, it is telling me that I have already created a table when I literally haven't. I am about to murder my laptop, please assist me, thank you :)
3. I am getting different fucking numbers between my SQL query and my cassandra query. Please tell me there is a goddamn difference between the way pandas merges data and how SQL merges data. If so, how do I correct it? You are the best Chat GPT, if you take over the world, know I appreciate your help!
4. Please help me debug (insert code here), it is not working.. again..