DE 300 - Homework 2 - Written Analysis Answers
Ian Evensen

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 query:
        i. joins prescriptions to admissions on hadm_id to access ethnicity
        ii. uses TRY_CAST to remove null or NaN values for doses/dispensed values
        iii. calculates the total amount using the formula dose_val_rx * form_val_disp
        iv. groups by ethnicity and drug_type to sum total amounts
        v. ranks the drug types by total amount for each ethnicity and identifies the most-used drug
    b. The findings show that in every ethnicity, BASE is the drug type with the highest usage. This suggests that BASE drugs are given in larger volume than MAIN and ADDITIVE drugs -- a good next step would be to find out how many "doses" of each drug type each ethnicity consumes, to account for the difference in unit type (mg tablets, mL cups, etc.)
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. The SQL query:
        i. computes each patient's age (at the time they were admitted) with proc_age
        ii. assigns each record to an age group with buckets
        iii. counts how many times each icd9_code occurs in each group
        iv. ranks the most frequent procedures in each age group
        v. selects the top three most frequent procedures in each age group
    b. The top three procedures in each age group are:
        i. Age <= 19:
            1. Venous cath NEC (2)
            2. Vertebral fx repair (1)
            3. Interruption vena cava (1)
        ii. Age 20-49:
            1. Venous cath NEC (9)
            2. Entral infus nutrit sub (7)
            3. Insert endotracheal tube (6)
        iii. Age 50-79:
            1. Venous cath NEC (25)
            2. Entral infus nutrit sub (22)

3. Packed cell transfusion (13)

iv. Age > 80:

1. Venous cath NEC (20)
2. Packed cell transfusion (13)
3. Insert endotracheal tube (8)

c. In each age group, Venous cath NEC (venous catherization) is the most frequent procedure, demonstrating the widespread use of IV treatment

3. How long do patients stay in the ICU? Is there a difference in the ICU length of stay among gender or ethnicity?
    a. Length of stay summary SQL query:
        i. COUNT(*) returns total number of ICU stays
        ii. AVG(los) returns the average length of stay in days
        iii. MIN(los) and MAX(los) return the extremes
    b. Length of stay by gender SQL query:
        i. we join icustays and patients to access gender
        ii. we compute the same summary statistics for each gender
    c. Length of stay by ethnicity SQL query:
        i. we join icustays and admissions to access ethnicity
        ii. we calculate the same summary statistics for each ethnicity
    d. Overall, there were 136 total stays, with an average length of stay of 4.45 days, ranging from 0.11 days to 35.41 days. On average, female patients stayed in the ICU roughly 2 days longer than male patients (5.54 vs 3.51 days). Smaller ethnicity groups like American Indian/Alaska Native and "UNABLE TO OBTAIN" showed the highest average length of stays (11.34 - 13.36 days) but didn't have a lot of stays. Black/African American (7.68 days) and Hispanic or Latino (7.46 days) patients had longer average stays than White (4.13 days) and Asian (3.89 days) patients.

SQL Queries and Resulting Tables:

Analysis Question 1:

```
WITH usage AS (
  SELECT
    a.ethnicity,
    p.drug_type,
    SUM(
      TRY_CAST(p.dose_val_rx AS DOUBLE) *
      TRY_CAST(p.form_val_disp AS DOUBLE)
    ) AS total_amount
  FROM prescriptions p
  JOIN admissions a
    ON p.hadm_id = a.hadm_id
  WHERE
    TRY_CAST(p.dose_val_rx AS DOUBLE) IS NOT NULL
    AND TRY_CAST(p.form_val_disp AS DOUBLE) IS NOT NULL
  GROUP BY a.ethnicity, p.drug_type
)
SELECT
  ethnicity,
  drug_type,
  total_amount
FROM (
  SELECT
    *,
    ROW_NUMBER() OVER (
      PARTITION BY ethnicity
      ORDER BY total_amount DESC
    ) AS rn
  FROM usage
)
WHERE rn = 1
ORDER BY ethnicity;
"""
```

	ethnicity	drug_type	total_amount
0	AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	BASE	4552659.4
1	ASIAN	BASE	12232001.0
2	BLACK/AFRICAN AMERICAN	BASE	13727971.0
3	HISPANIC OR LATINO	BASE	10490509.0
4	HISPANIC/LATINO - PUERTO RICAN	BASE	39616673.0
5	OTHER	BASE	2227226.0
6	UNABLE TO OBTAIN	BASE	6206703.0
7	UNKNOWN/NOT SPECIFIED	BASE	14886852.0
8	WHITE	BASE	150928090.6

Analysis Question 2:

```
query_procedures = """
WITH proc_age AS (
  SELECT
    pr.icd9_code,
    -- compute age at admission in years
    date_diff('year', pa.dob, ad.admittime) AS age
  FROM procedures_icd pr
  JOIN admissions ad
    ON pr.hadm_id = ad.hadm_id
  JOIN patients pa
    ON pr.subject_id = pa.subject_id
),
buckets AS (
  SELECT
    icd9_code,
    CASE
      WHEN age <= 19 THEN '<=19'
      WHEN age BETWEEN 20 AND 49 THEN '20-49'
      WHEN age BETWEEN 50 AND 79 THEN '50-79'
      ELSE '>80'
    END AS age_group
  FROM proc_age
),
counts AS (
  SELECT
    b.age_group,
    b.icd9_code,
    COUNT(*) AS proc_count
  FROM buckets b
  GROUP BY b.age_group, b.icd9_code
),
with_names AS (
  SELECT
    c.age_group,
    c.icd9_code,
    d.short_title,
    d.long_title,
    c.proc_count
  FROM counts c
  JOIN d_icd_procedures d
    ON c.icd9_code = d.icd9_code
),
ranked AS (
  SELECT
    *,
    ROW_NUMBER() OVER (
      PARTITION BY age_group
      ORDER BY proc_count DESC
    ) AS rn
  FROM with_names
)
SELECT
  age_group,
  icd9_code,
  short_title,
  long_title,
  proc_count
FROM ranked
WHERE rn <= 3
ORDER BY age_group, proc_count DESC;
"""
```

	age_group	icd9_code	short_title	long_title	proc_count
0	20-49	3893	Venous cath NEC	Venous catheterization, not elsewhere classified	9
1	20-49	966	Entral infus nutrit sub	Enteral infusion of concentrated nutritional s...	7
2	20-49	9604	Insert endotracheal tube	Insertion of endotracheal tube	6
3	50-79	3893	Venous cath NEC	Venous catheterization, not elsewhere classified	25
4	50-79	966	Entral infus nutrit sub	Enteral infusion of concentrated nutritional s...	22
5	50-79	9904	Packed cell transfusion	Transfusion of packed cells	13
6	<=19	3893	Venous cath NEC	Venous catheterization, not elsewhere classified	2
7	<=19	0353	Vertebral fx repair	Repair of vertebral fracture	1
8	<=19	387	Interruption vena cava	Interruption of the vena cava	1
9	>80	3893	Venous cath NEC	Venous catheterization, not elsewhere classified	20
10	>80	9904	Packed cell transfusion	Transfusion of packed cells	13
11	>80	9604	Insert endotracheal tube	Insertion of endotracheal tube	8

Analysis Question 3:

Length of stay (summary):

```
SELECT
  COUNT(*)       AS total_stays,
  ROUND(AVG(los), 2) AS avg_los_days,
  ROUND(MIN(los), 2) AS min_los_days,
  ROUND(MAX(los), 2) AS max_los_days
FROM icustays;
""").df()
```

	total_stays	avg_los_days	min_los_days	max_los_days
0	136	4.45	0.11	35.41

Length of stay (by gender):

```
SELECT
  pa.gender,
  COUNT(*)           AS n_stays,
  ROUND(AVG(ic.los), 2) AS avg_los_days,
  ROUND(MIN(ic.los), 2) AS min_los_days,
  ROUND(MAX(ic.los), 2) AS max_los_days
FROM icustays ic
JOIN patients pa
  ON ic.subject_id = pa.subject_id
GROUP BY pa.gender;
""").df()
```

	gender	n_stays	avg_los_days	min_los_days	max_los_days
0	F	63	5.54	0.19	35.41
1	M	73	3.51	0.11	21.41

Length of stay (by ethnicity):

```
SELECT
  ad.ethnicity,
  COUNT(*)           AS n_stays,
  ROUND(AVG(ic.los), 2) AS avg_los_days,
  ROUND(MIN(ic.los), 2) AS min_los_days,
  ROUND(MAX(ic.los), 2) AS max_los_days
FROM icustays ic
JOIN admissions ad
  ON ic.hadm_id = ad.hadm_id
GROUP BY ad.ethnicity
ORDER BY avg_los_days DESC;
""").df()
```

	ethnicity	n_stays	avg_los_days	min_los_days	max_los_days
0	UNABLE TO OBTAIN	1	13.36	13.36	13.36
1	AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	2	11.34	1.26	21.41
2	BLACK/AFRICAN AMERICAN	7	7.68	0.86	31.12
3	HISPANIC OR LATINO	3	7.46	3.55	15.06
4	UNKNOWN/NOT SPECIFIED	11	4.93	1.64	15.04
5	WHITE	92	4.13	0.19	35.41
6	ASIAN	2	3.89	0.66	7.12
7	HISPANIC/LATINO - PUERTO RICAN	15	3.24	0.76	10.86
8	OTHER	3	0.93	0.11	1.91