Cara Nugent 05/08/2025

Data Eng 300

Homework 02

GENAI STATEMENT: I did not use generative AI for this assignment

NOTE: All code is included in the google collab notebook with comments explaining.

PART ONE:

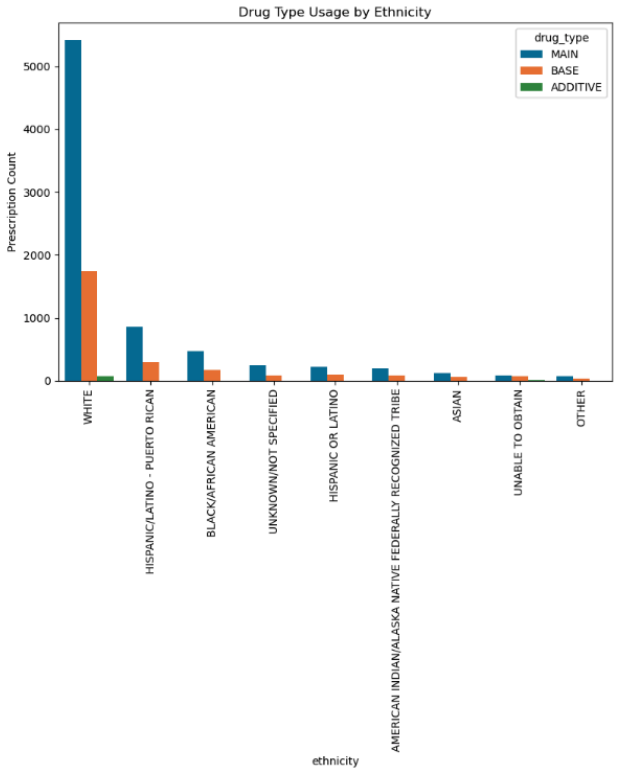
1. Drug type by ethnicity.
    a.
```
df1 = conn.sql(
    """
    SELECT
        A.ethnicity,
        PR.drug_type,
        COUNT(*) as total_use
    FROM PRESCRIPTIONS PR
    JOIN ADMISSIONS A ON PR.hadm_id = A.hadm_id
    WHERE drug_type IS NOT NULL
    GROUP BY ethnicity, drug_type
    ORDER BY total_use DESC
    """
).df()
```

    b. This query selects the ethnicity column from the admissions table, and drug_type column from the prescriptions table, joins them on their common hadm_id table, and creates a new column, count, that keeps track of the counts of each ethnicity and drug_type combination where there is a valid drug_type. The results are ordered by highest to lowest total_use.

    c.

	ethnicity	drug_type	total_use
0	WHITE	MAIN	5420
1	WHITE	BASE	1743
2	HISPANIC/LATINO - PUERTO RICAN	MAIN	860
3	BLACK/AFRICAN AMERICAN	MAIN	476
4	HISPANIC/LATINO - PUERTO RICAN	BASE	298

d. The top usage in each group (by ethnicity) is main.



2. Procedures by age group.

a/b.
```
conn.sql(
    """
    CREATE TABLE QUESTION02 AS
    SELECT
        d.short_title,
        a.admittime,
        p.dob,
        NULL AS age_group
    FROM procedures_icd PR
    JOIN d_icd_procedures D ON PR.icd9_code = D.icd9_code
    JOIN admissions A ON PR.hadm_id = A.hadm_id
    JOIN patients P ON A.subject_id = P.subject_id
    """)
```

This query creates a new table ‘QUESTION02’ using the short_title column from the d_icd_procedures table, admittime from admissions, dob from patients, and an empty age group column to be filled in later. It uses the procedures_icd table to join the d_icd_procedures table to the admissions table as d_icd_procedures only had the icd9_code identifier, and this was only in the procedures_idc table otherwise.

```
conn.sql(
    """
    ALTER TABLE QUESTION02 ALTER COLUMN age_group TYPE
TEXT;
""")
```

This query adds a new column ‘age_group’ of type text to my table.

```
conn.sql(
    """
    UPDATE QUESTION02
    SET age_group = CASE
        WHEN ROUND(ROUND(CAST(admittime AS DATE) -
CAST(dob AS DATE), 1) / 365.25) <= 19 THEN '<=19'
        WHEN ROUND(ROUND(CAST(admittime AS DATE) -
CAST(dob AS DATE), 1) / 365.25) BETWEEN 20 AND 49 THEN
'20-49'
        WHEN ROUND(ROUND(CAST(admittime AS DATE) -
CAST(dob AS DATE), 1) / 365.25) BETWEEN 50 AND 79 THEN
'50-79'
        WHEN ROUND(ROUND(CAST(admittime AS DATE) -
CAST(dob AS DATE), 1) / 365.25) >= 80 THEN '>=80'
        ELSE 'Unknown'
    END;
    """)
```

This query determines the age_group for each column by casting both the admittime and dobs to dates and subtracting them to find the days in between. Then that value is divided by 365.25 to find its value in years and the proper age_group title is added to that row.

```
df2 = conn.execute(
    """
    WITH procedures_by_age_group AS (
        SELECT age_group, short_title, COUNT(*) AS
procedure_count
        FROM QUESTION02
        GROUP BY age_group, short_title
    ),
    ranked AS (
        SELECT *, ROW_NUMBER() OVER (PARTITION BY
age_group ORDER BY procedure_count DESC) AS rank
        FROM procedures_by_age_group
    )
    SELECT age_group, short_title, procedure_count
    FROM ranked
```

```
    WHERE rank <= 3
    ORDER BY
        CASE age_group
            WHEN '<=19' THEN 1
            WHEN '20-49' THEN 2
            WHEN '50-79' THEN 3
            WHEN '>=80' THEN 4
            ELSE 6
        END,
        procedure_count DESC
""").df()
```

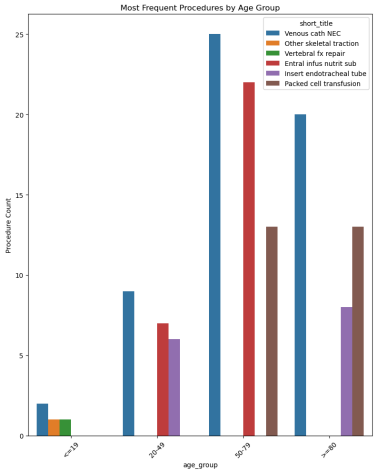
This query creates a subset of data ‘procedures_by_age_group’ from QUESTION02 and finds the rank of each procedure by age_group by partitioning the data by age_group and sorting by descending order of counts per procedure . Then the data is age_group, short_title, and procedure_count is grabbed from all of the rows with a rank of 3 or higher and the data ir ordered by age_group and procedure_count.

	age_group	short_title	procedure_count
0	<=19	Venous cath NEC	2
1	<=19	Other skeletal traction	1
2	<=19	Vertebral fx repair	1
3	20-49	Venous cath NEC	9
4	20-49	Entral infus nutrit sub	7

c.

d. The top 3 procedures reported for each age group are as follows:

- <=19: Venous cath NEC, Other skeletal traction, vertebral fx repair
- 20-49: Venous cath NEC, Entral infus nutrit sub, insert endotracheal tube
- 50-79:Venous cath NEC, Entral infus nutrit sub, packed cell transfusion
- >=80: Venous cath NEC, insert endotracheal tube, packed cell transfusion



3. ICU stays

a/b.
```
conn.sql(
    """
    CREATE TABLE QUESTION03 AS
    SELECT
        i.intime,
        i.outtime,
        a.ethnicity,
        p.gender,
        CAST(i.outtime AS DATE) - CAST(i.intime AS DATE)
as stay_time_days
    FROM icustays i
    JOIN admissions a ON i.hadm_id = a.hadm_id
    JOIN patients p ON i.subject_id = p.subject_id;
    """)
```

This query creates a new table ‘QUESTION03 using data from the icustays, admissions, and patients table. It also creates a column stay_time_days by subtracting the inttime in days from the outtime in days to get the total days stayed.

```
conn.execute(
    """
    SELECT
        AVG(stay_time_days) AS avg_stay,
        MAX(stay_time_days) - MIN(stay_time_days) AS
range_stay,
        STDDEV(stay_time_days) AS stddev_stay
    FROM QUESTION03
""")
```

This uses the data from QUESTION03 to return the average stay time, range of stay times, and standard deviation of stay times.

```
df3e = conn.execute(
    """
    SELECT ethnicity, stay_time_days, COUNT(*) as count
    FROM QUESTION03
    GROUP BY ethnicity, stay_time_days
    ORDER BY ethnicity, stay_time_days
    """).df()
```

This query groups the data by ethnicity and stay_times, counting the occurrences of each.

This uses the data from QUESTION03 to return the average stay time, range of stay times, and standard deviation of stay times.

```
df3g = conn.execute(
    """
    SELECT gender, stay_time_days, COUNT(*) as count
    FROM QUESTION03
    GROUP BY gender, stay_time_days
    ORDER BY gender, stay_time_days
    """).df()
```

This query groups the data by gender and stay_times, counting the occurrences of each.

c.

	avg_stay	range_stay	stddev_stay
0	4.426471	36	6.201114

(from second query)

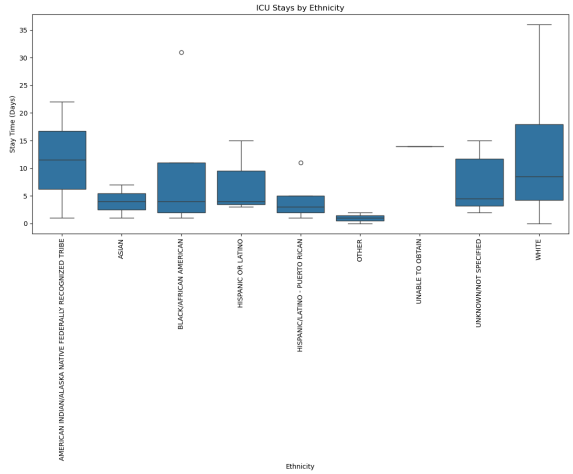
	ethnicity	stay_time_days	count
0	AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	1	1
1	AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	22	1
2	ASIAN	1	1
3	ASIAN	7	1
4	BLACK/AFRICAN AMERICAN	1	2

(third query)

	gender	stay_time_days	count
0	F	0	1
1	F	1	16
2	F	2	18
3	F	3	7
4	F	4	5

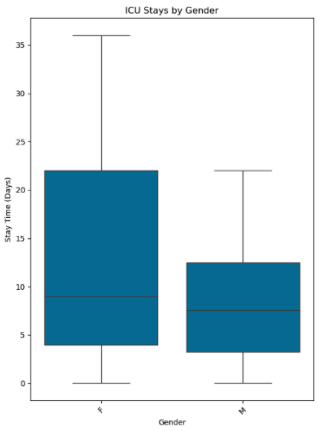
(fourth query)

d.



Based on the data, there is a difference in length of ICU stay by ethnicity. WHITE and AMERICAN INDIAN ethnicities tend to have the longest stays while OTHER appears to have the shortest stays. PUERTO RICAN and ASIAN ethnicities appear to have the shortest stays besides OTHER.

While the average ICU stay duration is similar between males and females (around 7–8 days), the distribution of stay times differs. Females have a noticeably higher upper quartile, indicating that the longer-stay cases are more common or more extreme in this group. Additionally, the maximum observed stay for females (36 days) is significantly higher than that for males (22 days), suggesting the presence of potential outliers. These outliers may be skewing the distribution for females, even though the average stay remains close to that of males.



PART TWO

NOTE: the information for parts A-C of all questions for part 2 are explicitly labeled a-c in the code. The code for d is labeled there as well but the verification can also be found below.

1.

```
Ethnicity: OTHER, Drug Type: BASE, Count: 28
Ethnicity: OTHER, Drug Type: MAIN, Count: 72
Ethnicity: BLACK/AFRICAN AMERICAN, Drug Type: BASE, Count: 169
Ethnicity: BLACK/AFRICAN AMERICAN, Drug Type: MAIN, Count: 476
Ethnicity: WHITE, Drug Type: ADDITIVE, Count: 66
Ethnicity: WHITE, Drug Type: BASE, Count: 1743
Ethnicity: WHITE, Drug Type: MAIN, Count: 5420
Ethnicity: ASIAN, Drug Type: BASE, Count: 56
Ethnicity: ASIAN, Drug Type: MAIN, Count: 121
Ethnicity: HISPANIC/LATINO - PUERTO RICAN, Drug Type: BASE, Count: 298
Ethnicity: HISPANIC/LATINO - PUERTO RICAN, Drug Type: MAIN, Count: 860
Ethnicity: UNKNOWN/NOT SPECIFIED, Drug Type: BASE, Count: 79
Ethnicity: UNKNOWN/NOT SPECIFIED, Drug Type: MAIN, Count: 245
Ethnicity: UNABLE TO OBTAIN, Drug Type: ADDITIVE, Count: 4
Ethnicity: UNABLE TO OBTAIN, Drug Type: BASE, Count: 68
Ethnicity: UNABLE TO OBTAIN, Drug Type: MAIN, Count: 89
Ethnicity: AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGNIZED TRIBE, Drug Type: ADDITIVE, Count: 2
Ethnicity: AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGNIZED TRIBE, Drug Type: BASE, Count: 80
Ethnicity: AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGNIZED TRIBE, Drug Type: MAIN, Count: 200
Ethnicity: HISPANIC OR LATINO, Drug Type: BASE, Count: 96
Ethnicity: HISPANIC OR LATINO, Drug Type: MAIN, Count: 226
```

Comparing the results from the cassandra query to the values returned by the table in part 1, and in the graph it is clear that the cassandra query produces the desired data.

2.

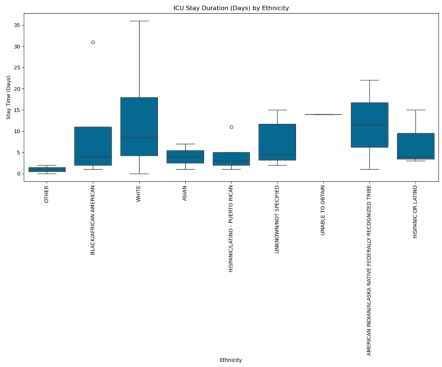
```
Age Group: 20-49, Procedure: Venous cath NEC, Count: 9
Age Group: 20-49, Procedure: Entral infus nutrit sub, Count: 7
Age Group: 20-49, Procedure: Cont inv mec ven 96+ hrs, Count: 6
Age Group: 50-79, Procedure: Venous cath NEC, Count: 25
Age Group: 50-79, Procedure: Entral infus nutrit sub, Count: 22
Age Group: 50-79, Procedure: Packed cell transfusion, Count: 13
Age Group: <=19, Procedure: Venous cath NEC, Count: 2
Age Group: <=19, Procedure: Applic ext fix dev-femur, Count: 1
Age Group: <=19, Procedure: Atlas-axis fusion, Count: 1
Age Group: >=80, Procedure: Venous cath NEC, Count: 20
Age Group: >=80, Procedure: Packed cell transfusion, Count: 13
Age Group: >=80, Procedure: Insert endotracheal tube, Count: 8
```

Comparing the results from the cassandra query to the values returned by the table in part 1, and in the graph it is clear that the cassandra query produces the desired data.

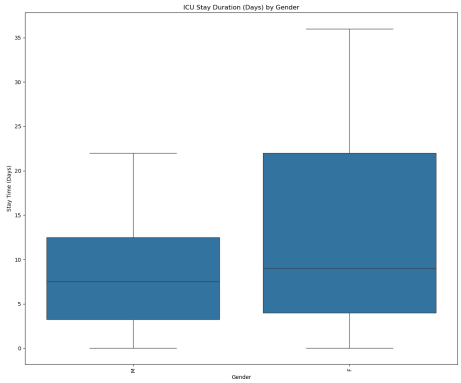
3.

```
Mean ICU Stay: 4.42647 days
Standard Deviation: 6.20111 days
Range: 36 days (min: 0, max: 36)
```

The mean, standard deviation, and range are the same as in part 1, so the correct data was extracted.



Comparing the data from the data extraction for stay duration by ethnicity, it is clear that this extracted the right data as the results appear the same as the results in part 1.



Comparing the data from the data extraction for stay duration by gender, it is clear that this extracted the right data as the results appear the same as the results in part 1.