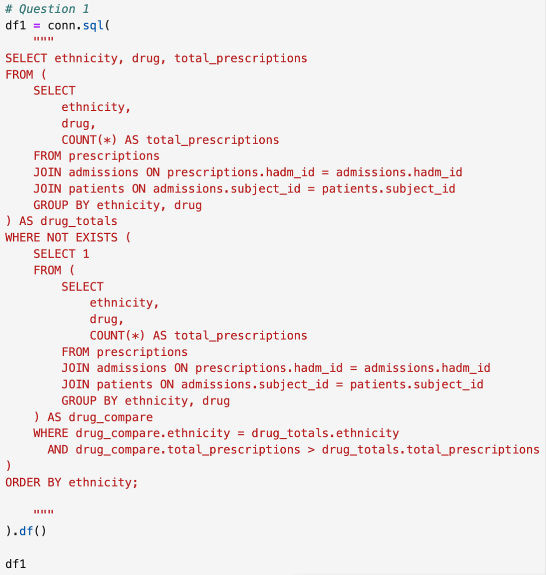
Part 1

Question 1

a.



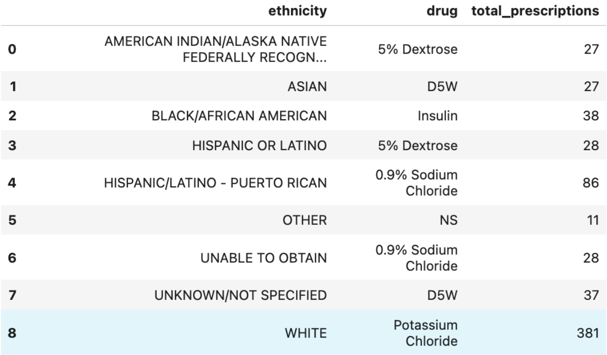
```
# Question 1
df1 = conn.sql(
    """
SELECT ethnicity, drug, total_prescriptions
FROM (
    SELECT
        ethnicity,
        drug,
        COUNT(*) AS total_prescriptions
    FROM prescriptions
    JOIN admissions ON prescriptions.hadm_id = admissions.hadm_id
    JOIN patients ON admissions.subject_id = patients.subject_id
    GROUP BY ethnicity, drug
) AS drug_totals
WHERE NOT EXISTS (
    SELECT 1
    FROM (
        SELECT
            ethnicity,
            drug,
            COUNT(*) AS total_prescriptions
        FROM prescriptions
        JOIN admissions ON prescriptions.hadm_id = admissions.hadm_id
        JOIN patients ON admissions.subject_id = patients.subject_id
        GROUP BY ethnicity, drug
    ) AS drug_compare
    WHERE drug_compare.ethnicity = drug_totals.ethnicity
      AND drug_compare.total_prescriptions > drug_totals.total_prescriptions
)
ORDER BY ethnicity;

    """
).df()

df1
```

b. This query joins prescriptions to patient ethnicity, groups by ethnicity and drug to count how often each drug was prescribed, and then filters to show only the top drug for each ethnicity based on total prescriptions.

c.

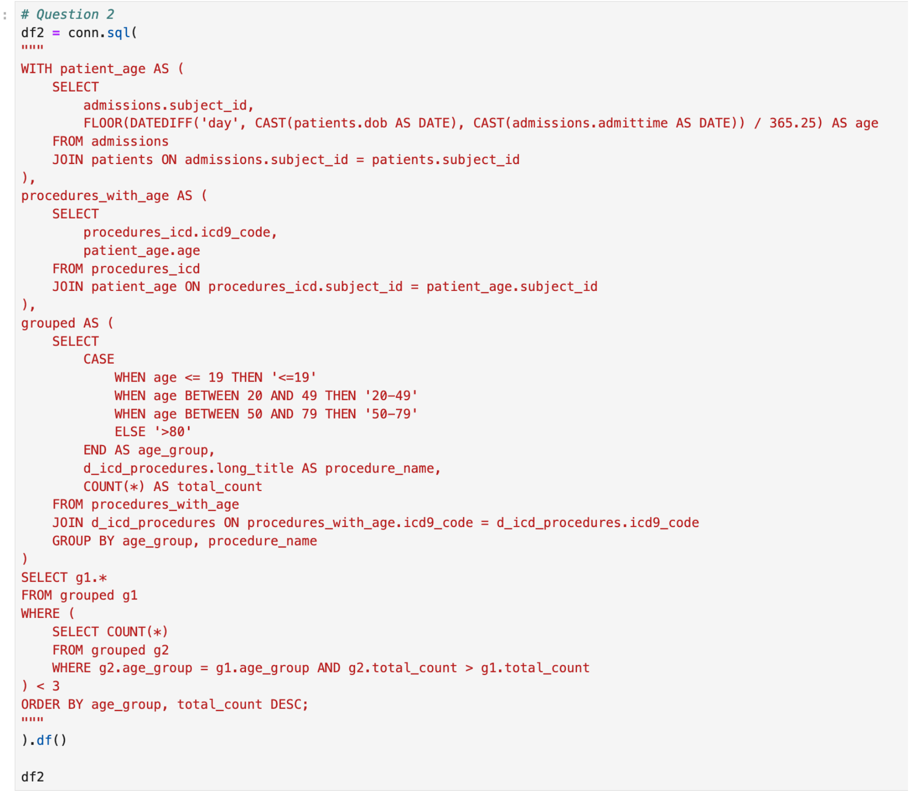


	ethnicity	drug	total_prescriptions
0	AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	5% Dextrose	27
1	ASIAN	D5W	27
2	BLACK/AFRICAN AMERICAN	Insulin	38
3	HISPANIC OR LATINO	5% Dextrose	28
4	HISPANIC/LATINO - PUERTO RICAN	0.9% Sodium Chloride	86
5	OTHER	NS	11
6	UNABLE TO OBTAIN	0.9% Sodium Chloride	28
7	UNKNOWN/NOT SPECIFIED	D5W	37
8	WHITE	Potassium Chloride	381

d. The top prescribed drug was different for each ethnicity, but there were some repeats. For example, Insulin was most common for Black/African American patients, and Potassium Chloride showed up the most overall, especially for White patients. A few drugs like 5% Dextrose, D5W, and 0.9% Sodium Chloride came up in multiple groups, which probably means they're just used a lot in general. Overall, the results show some differences between groups but also patterns in what's prescribed across the board.

Question 2

a.

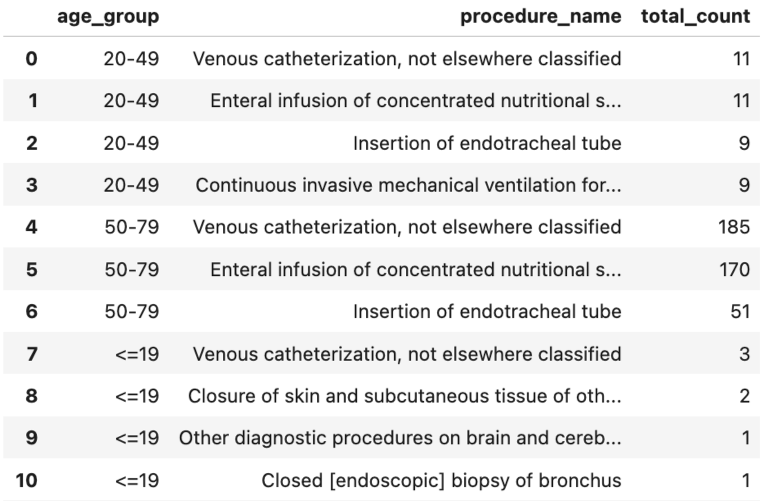


```
# Question 2
df2 = conn.sql(
"""
WITH patient_age AS (
    SELECT
        admissions.subject_id,
        FLOOR(DATEDIFF('day', CAST(patients.dob AS DATE), CAST(admissions.admittime AS DATE)) / 365.25) AS age
    FROM admissions
    JOIN patients ON admissions.subject_id = patients.subject_id
),
procedures_with_age AS (
    SELECT
        procedures_icd.icd9_code,
        patient_age.age
    FROM procedures_icd
    JOIN patient_age ON procedures_icd.subject_id = patient_age.subject_id
),
grouped AS (
    SELECT
        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,
        d_icd_procedures.long_title AS procedure_name,
        COUNT(*) AS total_count
    FROM procedures_with_age
    JOIN d_icd_procedures ON procedures_with_age.icd9_code = d_icd_procedures.icd9_code
    GROUP BY age_group, procedure_name
)
SELECT g1.*
FROM grouped g1
WHERE (
    SELECT COUNT(*)
    FROM grouped g2
    WHERE g2.age_group = g1.age_group AND g2.total_count > g1.total_count
) < 3
ORDER BY age_group, total_count DESC;
"""
).df()

df2
```

b. This query finds the top three most common procedures for each age group. I first calculated each patient's age at the time of admission and grouped them into age ranges. Then I joined that with the procedure data to count how often each procedure occurred in each group. Finally, I filtered the results to keep only the three procedures with the highest counts in each age group. It should be noted that if there are ties, they will be included and may extend outside three queries for an age group.

c.

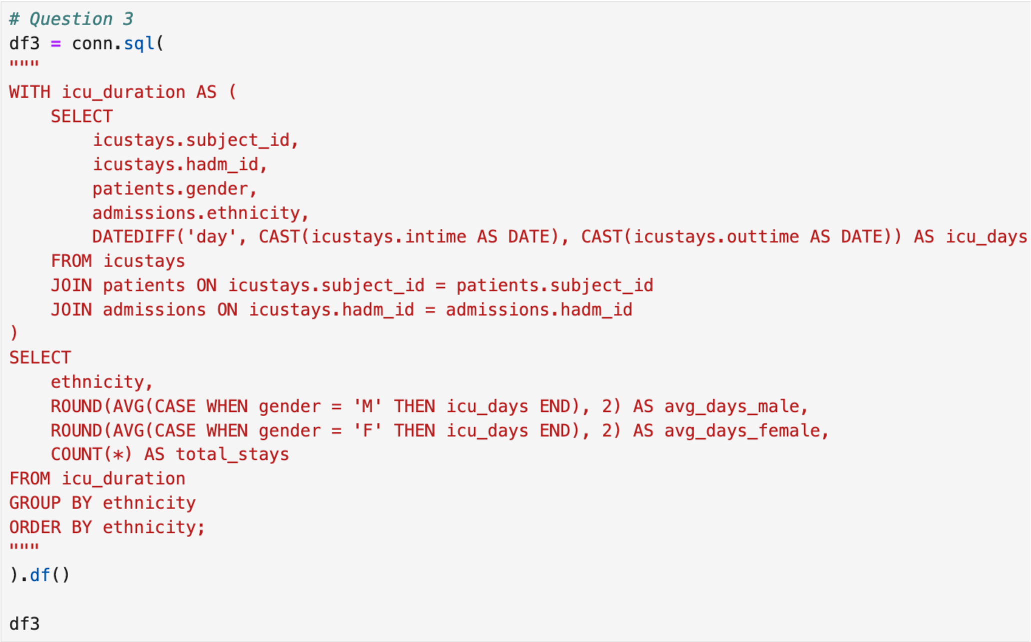


	age_group	procedure_name	total_count
0	20-49	Venous catheterization, not elsewhere classified	11
1	20-49	Enteral infusion of concentrated nutritional s...	11
2	20-49	Insertion of endotracheal tube	9
3	20-49	Continuous invasive mechanical ventilation for...	9
4	50-79	Venous catheterization, not elsewhere classified	185
5	50-79	Enteral infusion of concentrated nutritional s...	170
6	50-79	Insertion of endotracheal tube	51
7	<=19	Venous catheterization, not elsewhere classified	3
8	<=19	Closure of skin and subcutaneous tissue of oth...	2
9	<=19	Other diagnostic procedures on brain and cereb...	1
10	<=19	Closed [endoscopic] biopsy of bronchus	1

d. The most common procedures varied by age group. For patients aged 50–79, venous catheterization and nutritional infusions were the top procedures by far. The 20–49 group had a mix of similar procedures, including mechanical ventilation and endotracheal tube insertion. For those over 80, the top procedures were also fairly intensive, like catheterization and transfusions. Patients 19 and under had a wider variety of less frequent procedures, with no clear dominant one, which may reflect more specialized or varied cases in younger patients.

Question 3

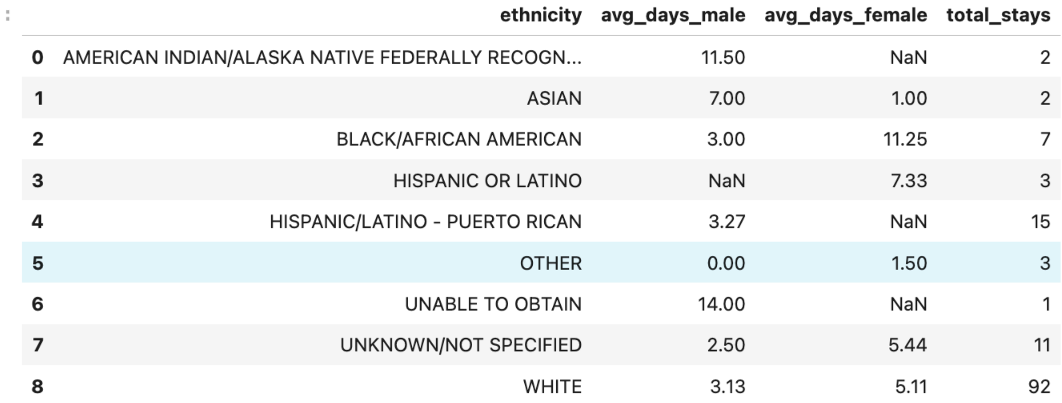
a.



```
# Question 3
df3 = conn.sql(
"""
WITH icu_duration AS (
    SELECT
        icustays.subject_id,
        icustays.hadm_id,
        patients.gender,
        admissions.ethnicity,
        DATEDIFF('day', CAST(icustays.intime AS DATE), CAST(icustays.outtime AS DATE)) AS icu_days
    FROM icustays
    JOIN patients ON icustays.subject_id = patients.subject_id
    JOIN admissions ON icustays.hadm_id = admissions.hadm_id
)
SELECT
    ethnicity,
    ROUND(AVG(CASE WHEN gender = 'M' THEN icu_days END), 2) AS avg_days_male,
    ROUND(AVG(CASE WHEN gender = 'F' THEN icu_days END), 2) AS avg_days_female,
    COUNT(*) AS total_stays
FROM icu_duration
GROUP BY ethnicity
ORDER BY ethnicity;
"""
).df()

df3
```

b. This query looks at how long patients stay in the ICU and compares the average stay between males and females within each ethnicity group. I calculated the ICU stay in days, then grouped the results by ethnicity while showing the average stay separately for male and female patients. This lets us see if there are noticeable differences in ICU length of stay based on both gender and ethnicity.



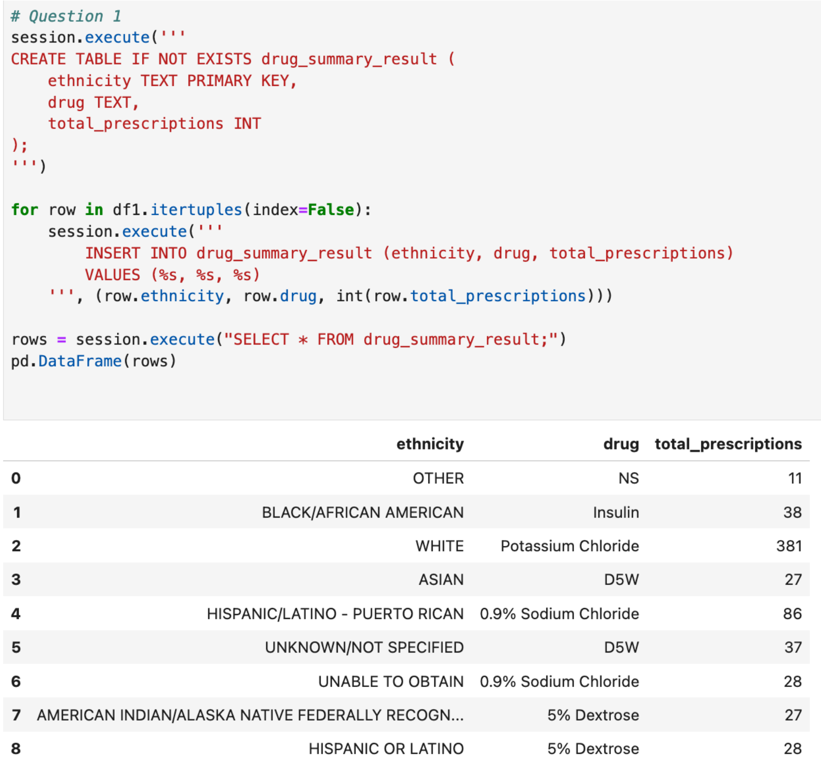
	ethnicity	avg_days_male	avg_days_female	total_stays
0	AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	11.50	NaN	2
1	ASIAN	7.00	1.00	2
2	BLACK/AFRICAN AMERICAN	3.00	11.25	7
3	HISPANIC OR LATINO	NaN	7.33	3
4	HISPANIC/LATINO - PUERTO RICAN	3.27	NaN	15
5	OTHER	0.00	1.50	3
6	UNABLE TO OBTAIN	14.00	NaN	1
7	UNKNOWN/NOT SPECIFIED	2.50	5.44	11
8	WHITE	3.13	5.11	92

c.

d. ICU length of stay varies by both gender and ethnicity, but some groups had very few stays, which affects the averages. For example, American Indian/Alaska Native patients had the highest male average at 11.5 days, while Black/African American females averaged 11.25 days. In contrast, many groups like Asian and Other had very short stays or missing values for one gender, likely due to low counts. White patients made up the largest group, with moderate ICU stays for both males (3.13 days) and females (5.11 days). Overall, some gender-ethnicity combinations stayed longer than others, but small sample sizes in many groups make it hard to draw strong conclusions.

Part 2

Question 1



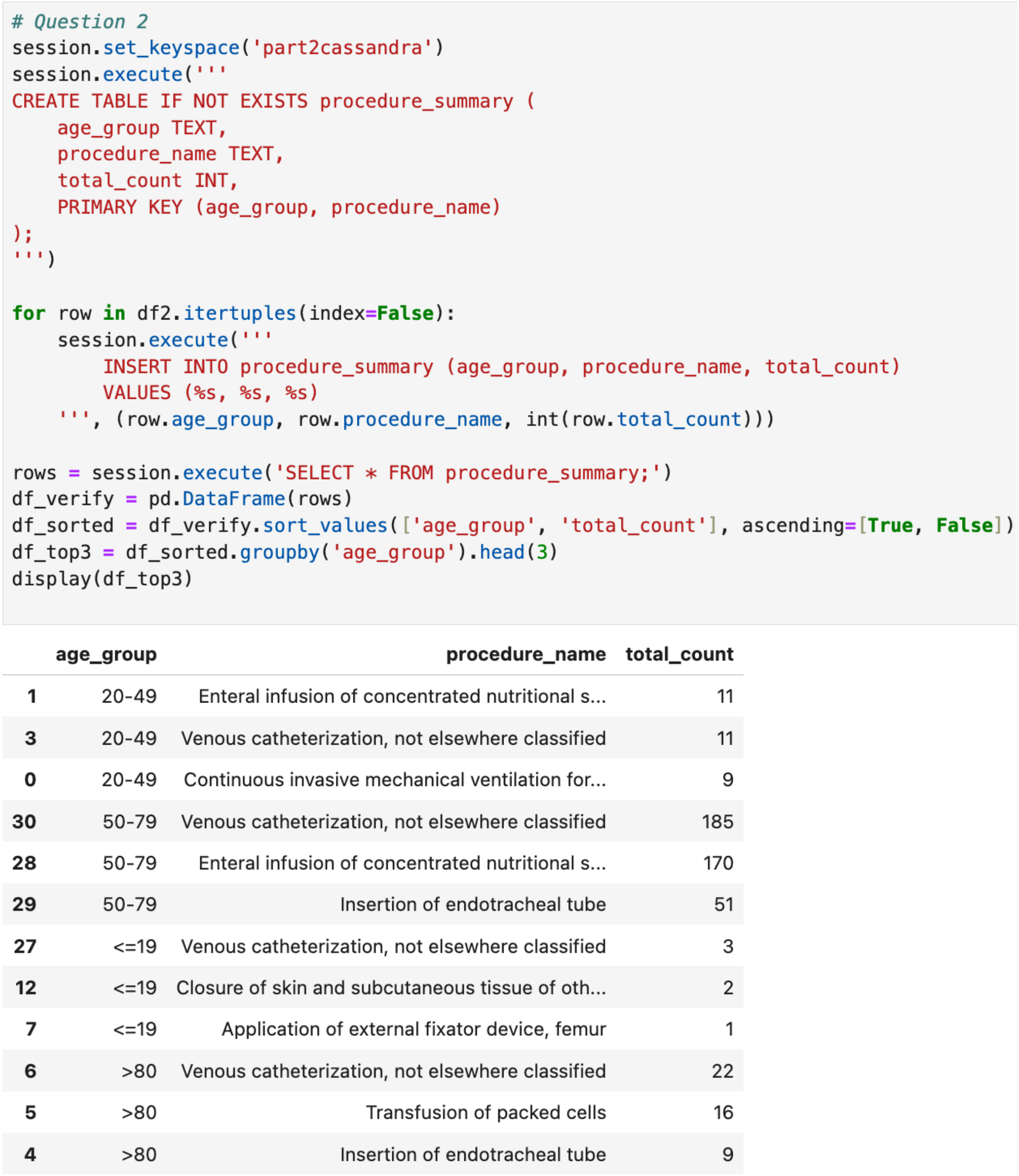
```
# Question 1
session.execute('''
CREATE TABLE IF NOT EXISTS drug_summary_result (
    ethnicity TEXT PRIMARY KEY,
    drug TEXT,
    total_prescriptions INT
);
''')

for row in df1.itertuples(index=False):
    session.execute('''
        INSERT INTO drug_summary_result (ethnicity, drug, total_prescriptions)
        VALUES (%s, %s, %s)
    ''', (row.ethnicity, row.drug, int(row.total_prescriptions)))

rows = session.execute("SELECT * FROM drug_summary_result;")
pd.DataFrame(rows)
```

	ethnicity	drug	total_prescriptions
0	OTHER	NS	11
1	BLACK/AFRICAN AMERICAN	Insulin	38
2	WHITE	Potassium Chloride	381
3	ASIAN	D5W	27
4	HISPANIC/LATINO - PUERTO RICAN	0.9% Sodium Chloride	86
5	UNKNOWN/NOT SPECIFIED	D5W	37
6	UNABLE TO OBTAIN	0.9% Sodium Chloride	28
7	AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	5% Dextrose	27
8	HISPANIC OR LATINO	5% Dextrose	28

Question 2



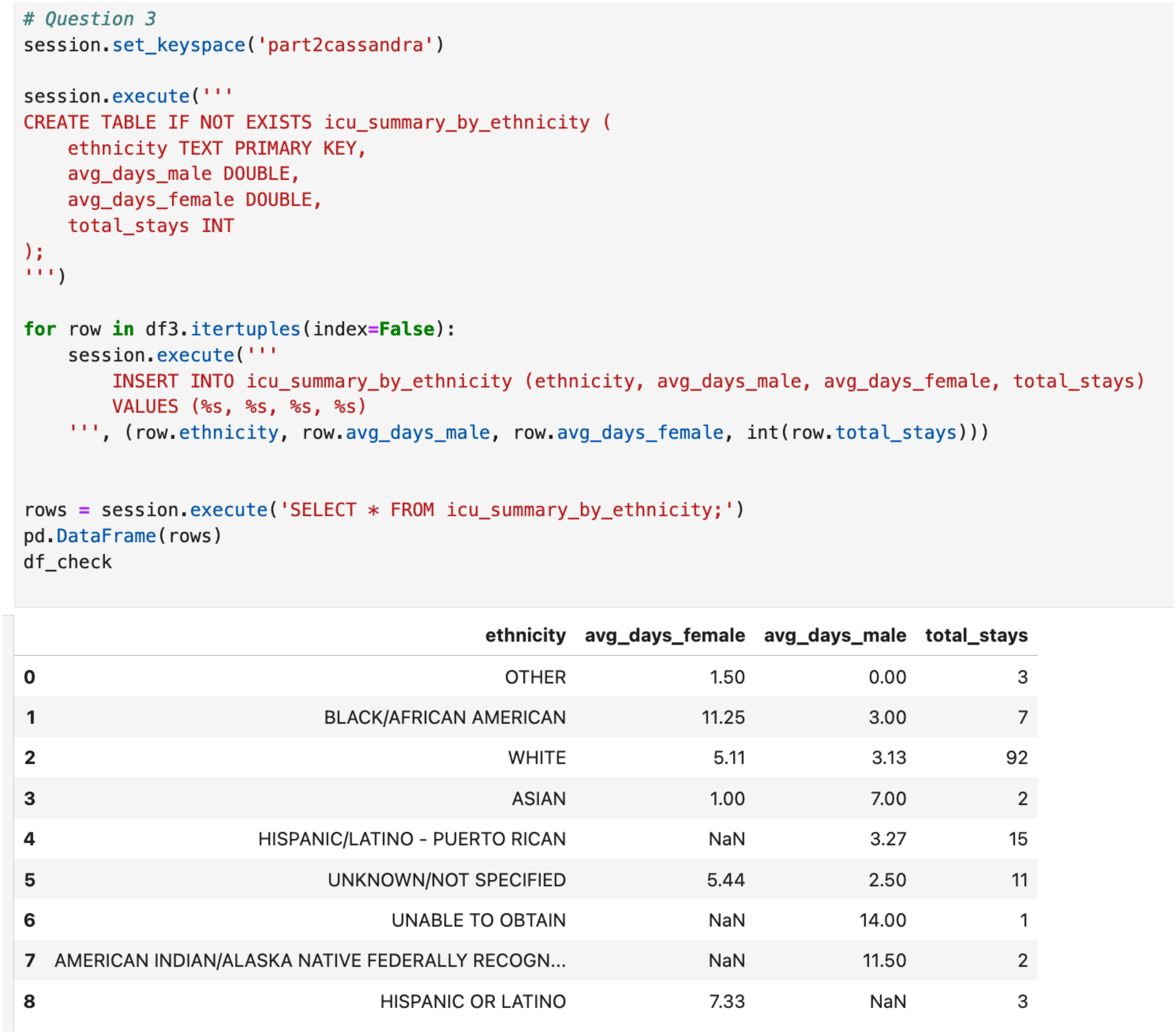
```
# Question 2
session.set_keyspace('part2cassandra')
session.execute('''
CREATE TABLE IF NOT EXISTS procedure_summary (
    age_group TEXT,
    procedure_name TEXT,
    total_count INT,
    PRIMARY KEY (age_group, procedure_name)
);
''')

for row in df2.itertuples(index=False):
    session.execute('''
        INSERT INTO procedure_summary (age_group, procedure_name, total_count)
        VALUES (%s, %s, %s)
    ''', (row.age_group, row.procedure_name, int(row.total_count)))

rows = session.execute('SELECT * FROM procedure_summary;')
df_verify = pd.DataFrame(rows)
df_sorted = df_verify.sort_values(['age_group', 'total_count'], ascending=[True, False])
df_top3 = df_sorted.groupby('age_group').head(3)
display(df_top3)
```

	age_group	procedure_name	total_count
1	20-49	Enteral infusion of concentrated nutritional s...	11
3	20-49	Venous catheterization, not elsewhere classified	11
0	20-49	Continuous invasive mechanical ventilation for...	9
30	50-79	Venous catheterization, not elsewhere classified	185
28	50-79	Enteral infusion of concentrated nutritional s...	170
29	50-79	Insertion of endotracheal tube	51
27	<=19	Venous catheterization, not elsewhere classified	3
12	<=19	Closure of skin and subcutaneous tissue of oth...	2
7	<=19	Application of external fixator device, femur	1
6	>80	Venous catheterization, not elsewhere classified	22
5	>80	Transfusion of packed cells	16
4	>80	Insertion of endotracheal tube	9

Question 3



```
# Question 3
session.set_keyspace('part2cassandra')

session.execute('''
CREATE TABLE IF NOT EXISTS icu_summary_by_ethnicity (
    ethnicity TEXT PRIMARY KEY,
    avg_days_male DOUBLE,
    avg_days_female DOUBLE,
    total_stays INT
);
''')

for row in df3.itertuples(index=False):
    session.execute('''
        INSERT INTO icu_summary_by_ethnicity (ethnicity, avg_days_male, avg_days_female, total_stays)
        VALUES (%s, %s, %s, %s)
    ''', (row.ethnicity, row.avg_days_male, row.avg_days_female, int(row.total_stays)))


rows = session.execute('SELECT * FROM icu_summary_by_ethnicity;')
pd.DataFrame(rows)
df_check
```

	ethnicity	avg_days_female	avg_days_male	total_stays
0	OTHER	1.50	0.00	3
1	BLACK/AFRICAN AMERICAN	11.25	3.00	7
2	WHITE	5.11	3.13	92
3	ASIAN	1.00	7.00	2
4	HISPANIC/LATINO - PUERTO RICAN	NaN	3.27	15
5	UNKNOWN/NOT SPECIFIED	5.44	2.50	11
6	UNABLE TO OBTAIN	NaN	14.00	1
7	AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGN...	NaN	11.50	2
8	HISPANIC OR LATINO	7.33	NaN	3