SQL Mini Project

1. Introduction

This SQL mini project focuses on analyzing 10,000 real-world medical appointment records from Virginia to understand the factors affecting patient attendance. The goal is to uncover patterns related to no-shows by examining demographic data (age, gender), scheduling details (day of the week, lead time), and clinical factors (medical conditions).

Key objectives include:

- Identifying trends that lead to missed appointments
- Improving patient communication strategies
- Reducing no-show rates
- Optimizing clinic scheduling and resource use

By using SQL for data exploration and analysis, this project supports more informed decision-making in healthcare operations.

2. Dataset Overview

This project uses a real-world dataset containing 10,000 medical appointment records from the state of Virginia, USA. Each row in the dataset represents a unique medical appointment and includes both demographic and clinical details about the patient, as well as whether or not they attended the appointment.

Dataset Shape

- **Rows:** 10,000 (each representing a single appointment)
- **Columns:** Multiple fields capturing patient details, appointment scheduling, and health indicators

Key Variables:

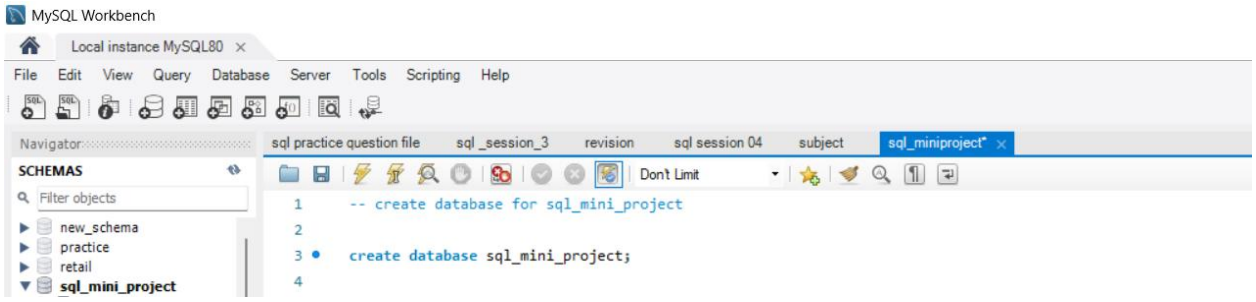
- PatientId: Unique identifier for each patient
- AppointmentID: Unique identifier for each appointment
- Gender: Gender of the patient
- Scheduled Day: Date and time when the appointment was scheduled

- Appointment Day: Date when the appointment was to take place
- Age: Age of the patient
- Neighbourhood: Area where the patient lives
- Scholarship: Indicates whether the patient is enrolled in a welfare program
- Hipertension, Diabetes, Alcoholism, Handcap: Indicators of health conditions
- SMS_received: Whether the patient received an SMS reminder
- Showed_up: Whether the patient attended the appointment
- Date.diff: Calculated difference in days between the scheduled and appointment date

3. Methodology and SQL Tasks

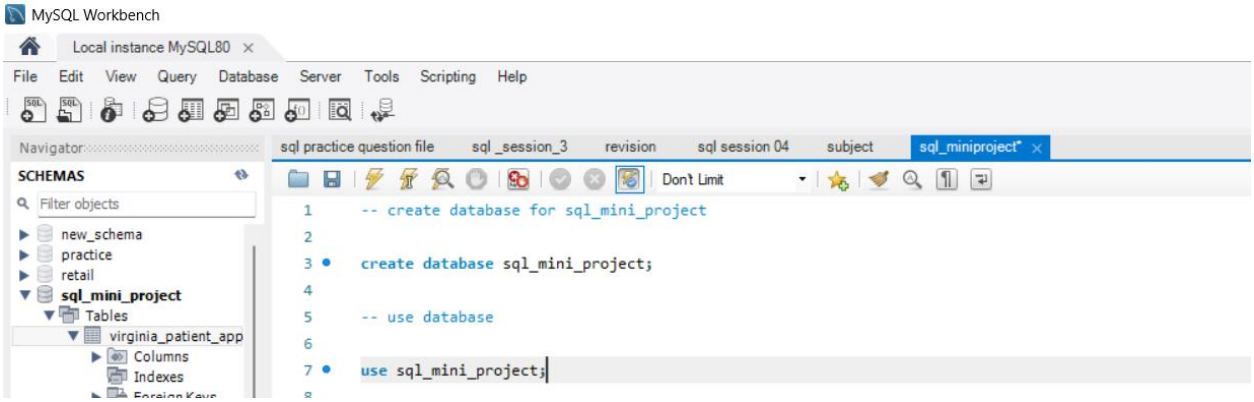
- **Create database for SQL mini project**

```
-- create database for sql_mini_project
create database sql_mini_project;
```

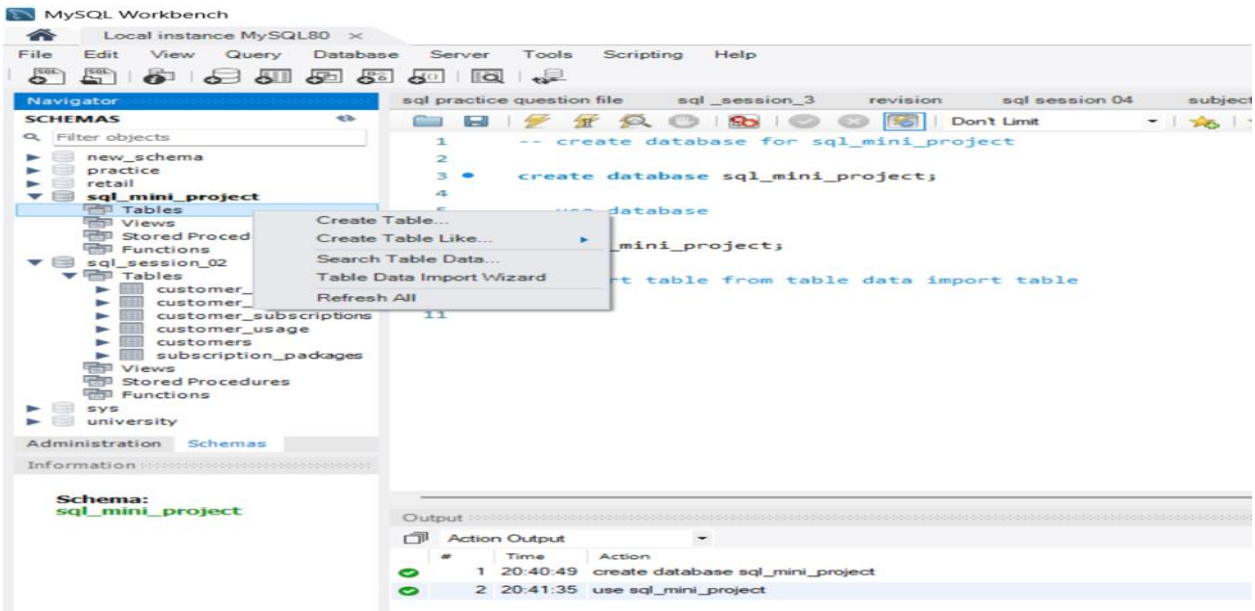


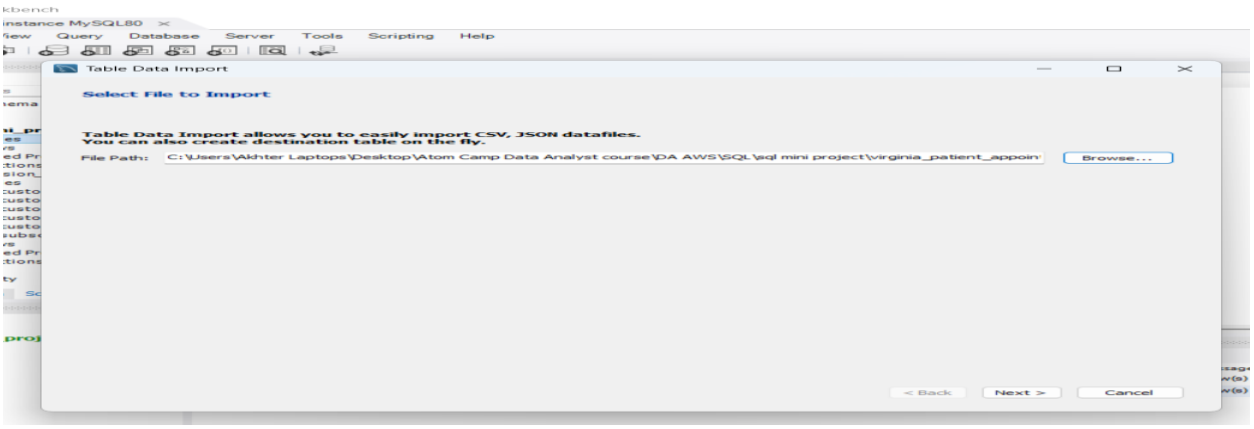
- **Use the create database for SQL mini project**

```
-- use database

use sql_mini_project;
```



- **Import the dataset for SQL mini project**





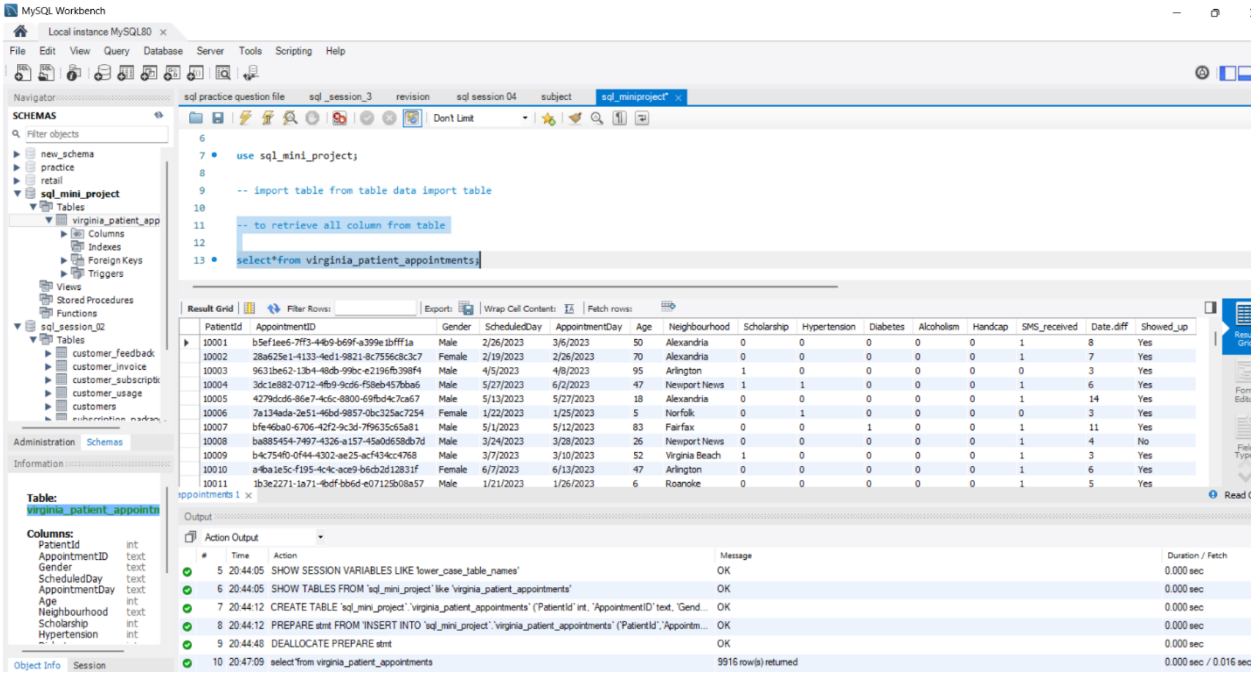
Summary / Interpretation

The SQL database for the project has been successfully created and loaded with the appointment dataset. It is now ready for analysis using SQL queries to explore patterns and insights related to patient attendance.

Basic SQL & Data Retrieval

1. Retrieve all columns from the Appointments table.

```
-- to retrieve all column from table
select*from virginia_patient_appointments;
```

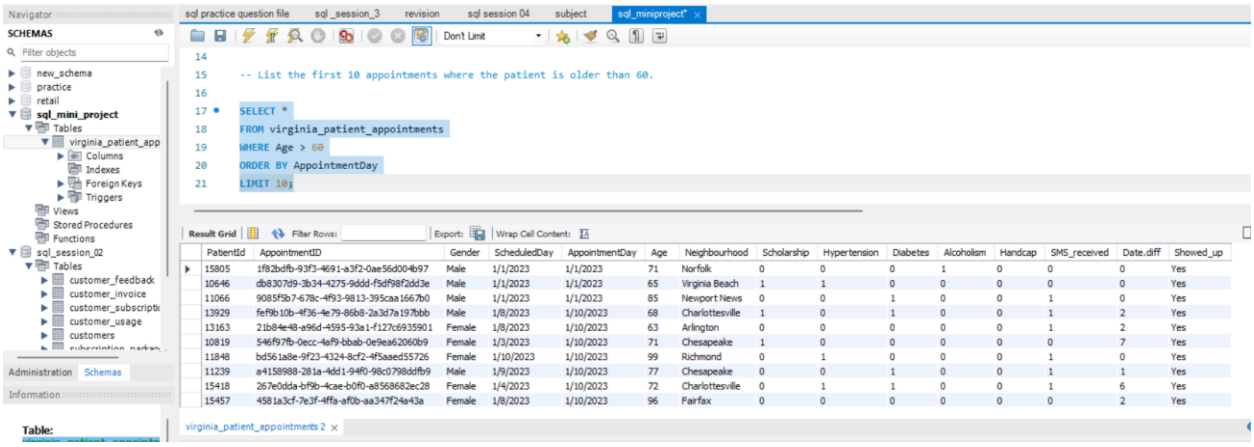


Summary / Interpretation

All columns from the Appointments table were successfully retrieved, providing a complete view of the dataset. This step helps confirm that the data is loaded correctly and is ready for analysis.

2. List the first 10 appointments where the patient is older than 60.

```
Select *
from virginia_patient_appointments
where Age > 60
order by AppointmentDay
limit 10;
```



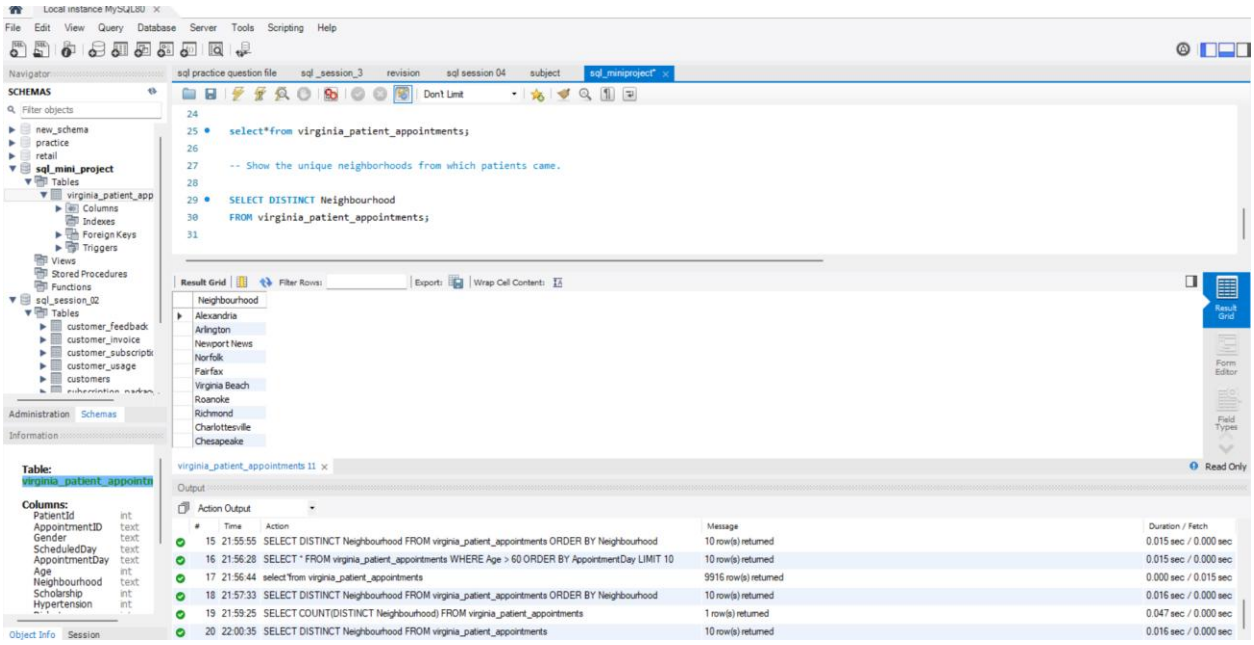
PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hypertension	Diabetes	Alcoholism	Handcap	SMS_received	Date.diff	Showed_up
15805	1f82bdfb-93f3-4691-a3f2-0ae56d004b97	Male	1/1/2023	1/1/2023	71	Norfolk	0	0	0	1	0	0	0	Yes
10646	db8307d9-3b34-4275-9ddd-f5df98f2dd3e	Male	1/1/2023	1/1/2023	65	Virginia Beach	1	1	0	0	0	0	0	Yes
11066	9085f5b7-678c-4f93-9813-395caa1667b0	Male	1/1/2023	1/1/2023	85	Newport News	0	0	1	0	0	1	0	Yes
13929	fef9b10b-4f36-4e79-86b8-2a3d7a197bbb	Male	1/8/2023	1/10/2023	68	Charlottesville	1	0	1	0	0	1	2	Yes
13163	21b84e48-a96d-4595-93a1-f127c6935901	Female	1/8/2023	1/10/2023	63	Arlington	0	0	0	0	0	1	2	Yes
10819	546f97fb-0ecc-4af9-bbab-0e9ea62060b9	Female	1/3/2023	1/10/2023	71	Chesapeake	1	0	0	0	0	0	7	Yes
11848	bd561a8e-9f23-4324-8cf2-4f5aaed55726	Female	1/10/2023	1/10/2023	99	Richmond	0	1	0	0	0	1	0	Yes
11239	a4158988-281a-4dd1-94f0-98c0798ddfb9	Male	1/9/2023	1/10/2023	77	Chesapeake	0	0	1	0	0	1	1	Yes
15418	267e0dda-bf9b-4cae-b0f0-a8568682ec28	Female	1/4/2023	1/10/2023	72	Charlottesville	0	1	1	0	0	1	6	Yes
15457	4581a3cf-7e3f-4ffa-af0b-aa347f24a43a	Female	1/8/2023	1/10/2023	96	Fairfax	0	0	0	0	0	0	2	Yes

Summary/Interpretation

The first 10 appointments for patients over 60 show a mix of genders and locations. All patients attended their appointments, suggesting good adherence among older individuals.

3. Show the unique neighborhoods from which patients came.

```
-- Show the unique neighborhoods from which patients came.
SELECT DISTINCT Neighbourhood
FROM virginia_patient_appointments;
```



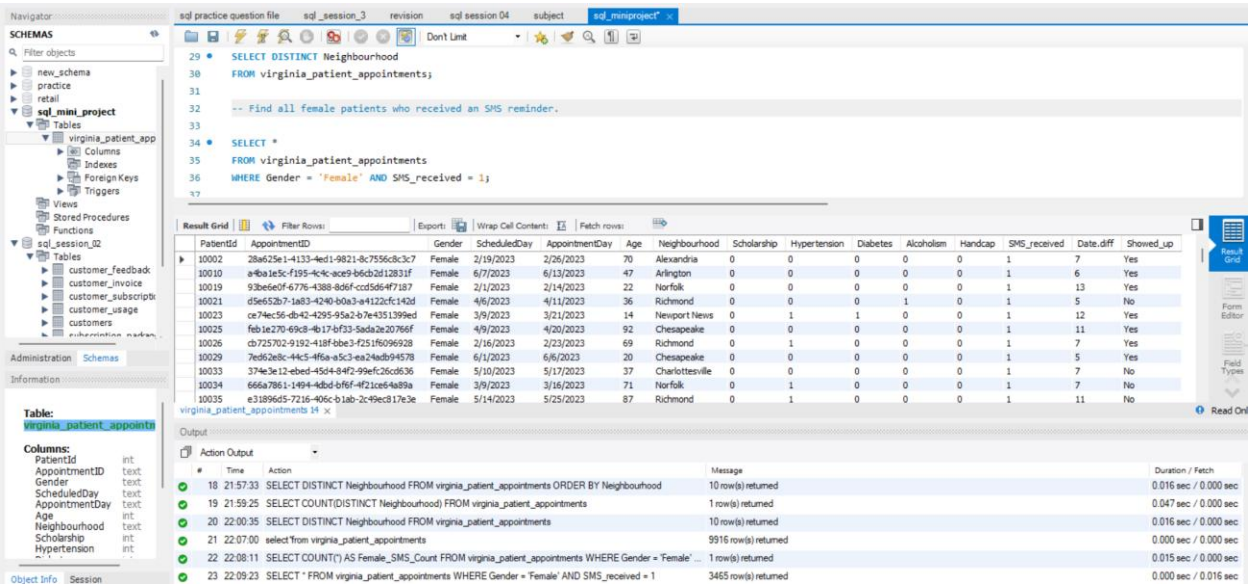
Summary / Interpretation

The dataset shows that patients come from various cities across Virginia, including Richmond, Norfolk, Virginia Beach, and Charlottesville. This indicates that the clinic serves a diverse and widespread population.

4. Find all female patients who received an SMS reminder. Give count of them

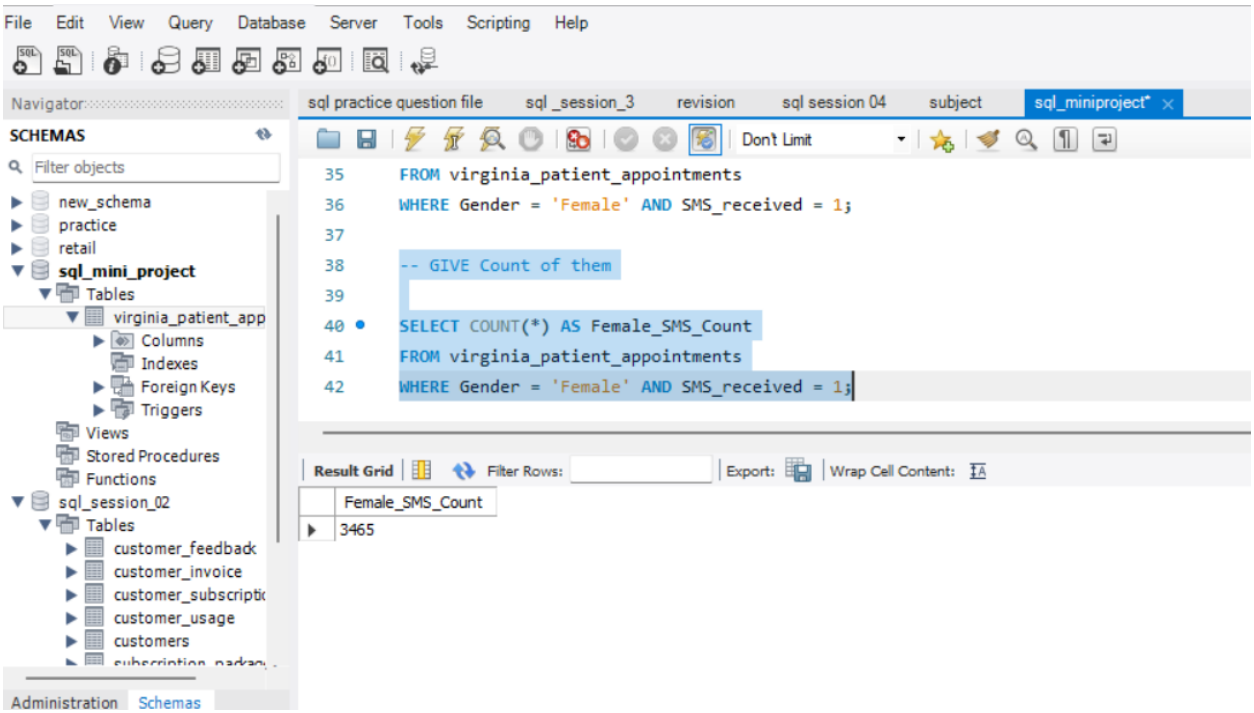
-- Find all female patients who received an SMS reminder.

```
SELECT *
FROM virginia_patient_appointments
WHERE Gender = 'Female' AND SMS_received = 1;
```



-- GIVE Count of them

```
SELECT COUNT (*) AS Female_SMS_Count
FROM virginia_patient_appointments
WHERE Gender = 'Female' AND SMS_received = 1;
```



Summary / Interpretation

A total of **3,465 female patients** received SMS reminders for their appointments. This indicates that SMS is actively used as a communication method for female patients.

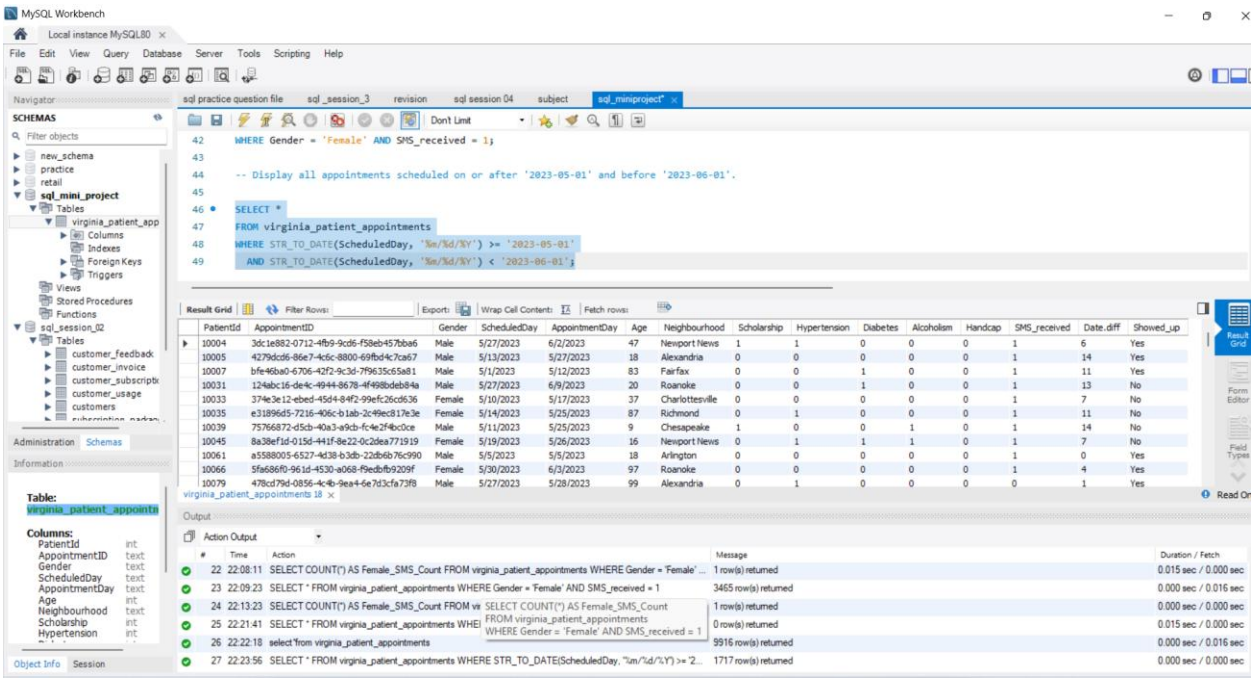
5. Display all appointments scheduled on or after '2023-05-01' and before '2023-06-01'.

SELECT *

FROM virginia_patient_appointments

WHERE STR_TO_DATE (ScheduledDay, '%m/%d/%Y') >= '2023-05-01'

AND STR_TO_DATE (ScheduledDay, '%m/%d/%Y') < '2023-06-01';



Summary / Interpretation

A total of 1,717 appointments were scheduled between May 1, 2023, and May 31, 2023. This dataset provides a snapshot of appointments within that specific time frame.

Data Modification & Filtering

6. Update the 'Showed_up' status to 'Yes' where it is null or empty

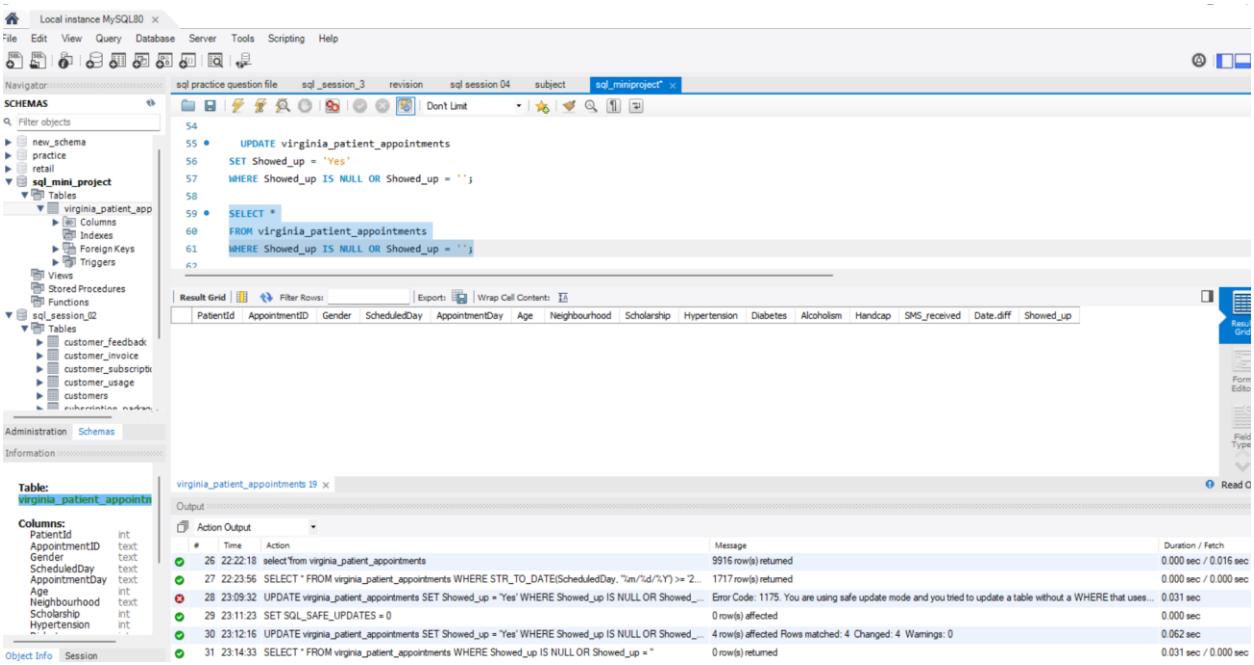
```
SET SQL_SAFE_UPDATES = 0;

UPDATE virginia_patient_appointments

SET Showed_up = 'Yes'

WHERE showed_up IS NULL OR Showed_up = ' ';
```

For check

```
SELECT *
```

FROM virginia_patient_appointments

WHERE showed_up IS NULL OR Showed_up = '';



Summary / Interpretation

The Showed_up status was successfully updated for 4 records where the value was previously NULL or empty.

7. Add a new column Appointment Status using a CASE statement: 'No Show' if showed_up = 'No' 'Attended' otherwise

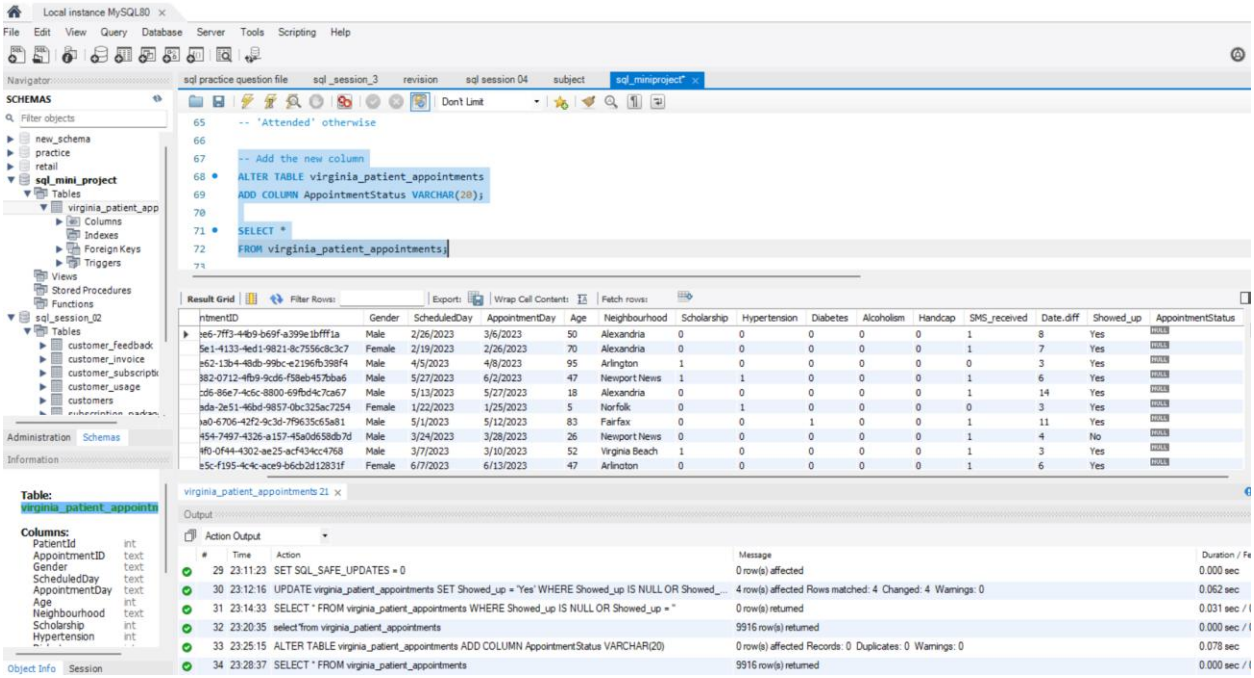
<u>-- Add the new column</u>

ALTER TABLE virginia_patient_appointments

ADD COLUMN AppointmentStatus VARCHAR (20);

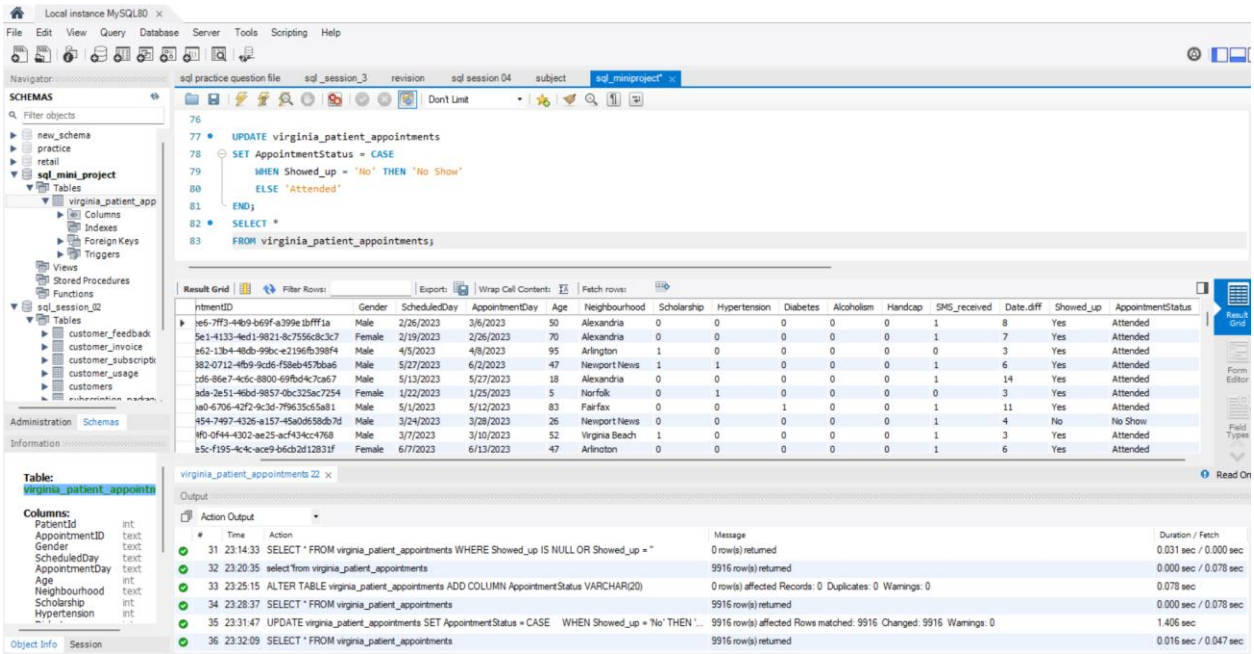
SELECT *

FROM virginia_patient_appointments;



<u>-- Update the column using CASE</u>

```
UPDATE virginia_patient_appointments
SET AppointmentStatus = CASE
 WHEN showed_up = 'No' THEN 'No Show'
   ELSE 'Attended'
END;
SELECT *
FROM virginia_patient_appointments;
```

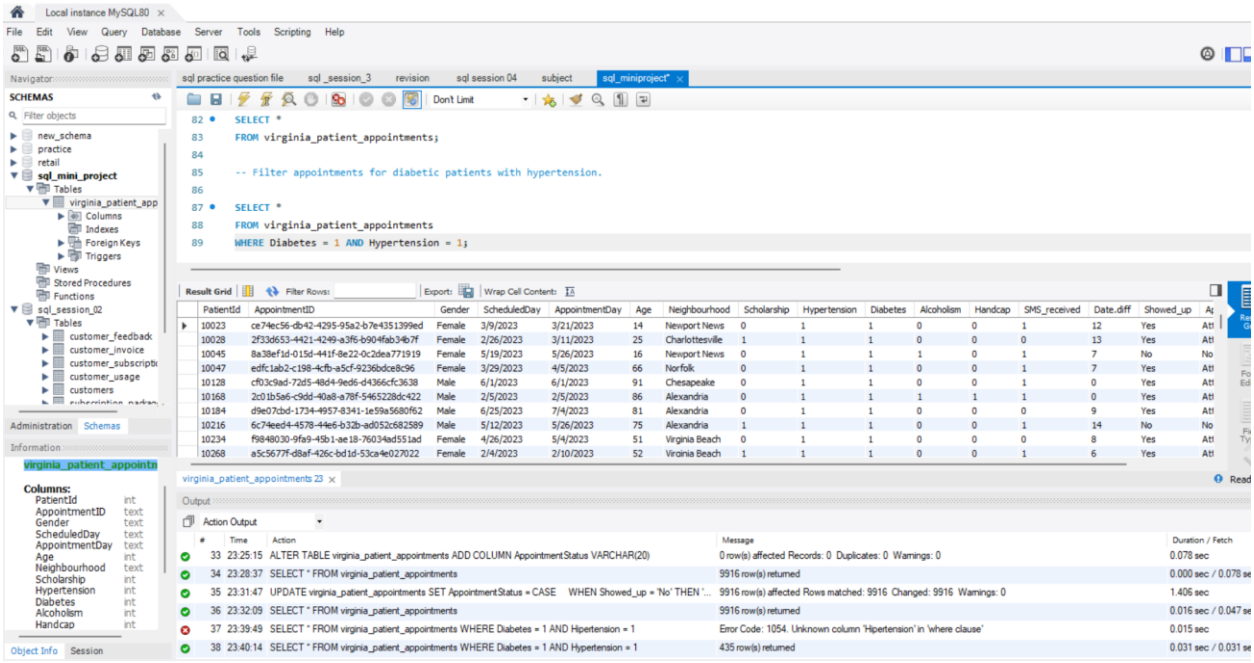


Summary / Interpretation

The AppointmentStatus column was successfully added and updated for 9,916 records based on the Showed_up status. While the exact counts of 'No Show' vs. 'Attended' were not calculated, the majority of records were filled with 'Attended', indicating that most patients showed up for their appointments.

8 Filter appointments for diabetic patients with hypertension.

```
SELECT *
FROM virginia_patient_appointments
WHERE Diabetes = 1 AND Hypertension = 1;
```

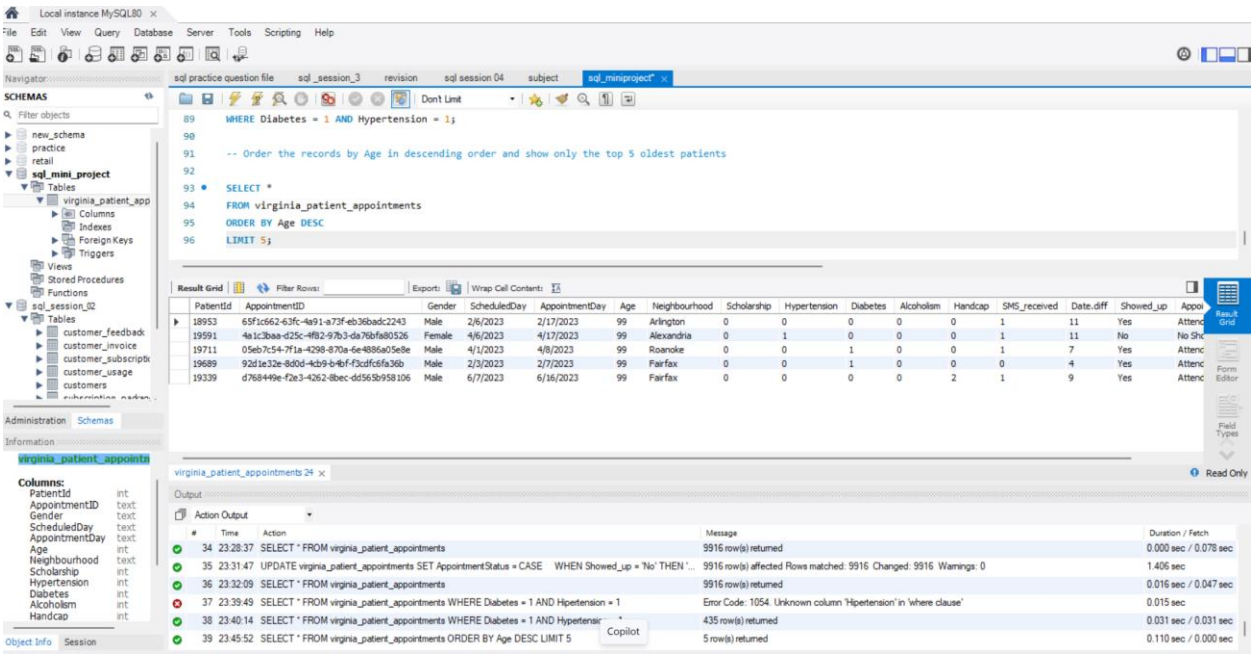


Summary / Interpretation

A total of 435 appointments were scheduled for patients who have both diabetes and hypertension. This subset of patients may require specialized care and attention, which could be useful for improving treatment plans and focusing on high-risk groups.

9 Order the records by Age in descending order and show only the top 5 oldest patients.

```
SELECT *
FROM virginia_patient_appointments
ORDER BY Age DESC
 LIMIT 5;
```

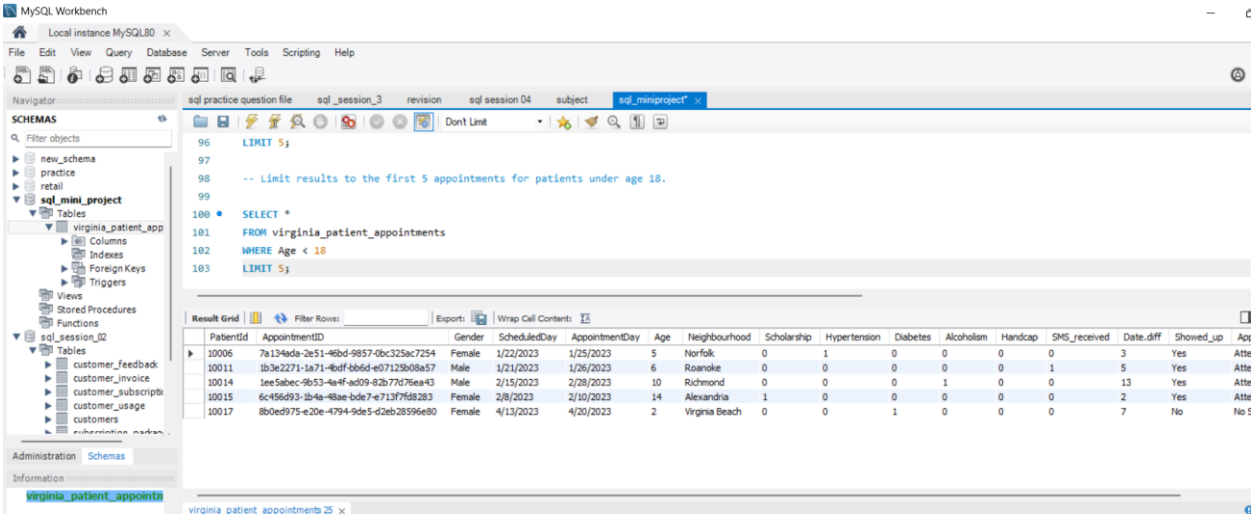


Summary / Interpretation

The 5 oldest patients in the dataset are all 99 years old, with 4 males and 1 female. This highlights the presence of elderly patients in the clinic, potentially requiring additional care and attention for age-related health concerns.

10 Limit results to the first 5 appointments for patients under age 18.

```
SELECT *
FROM virginia_patient_appointments
WHERE Age < 18
LIMIT 5;
```



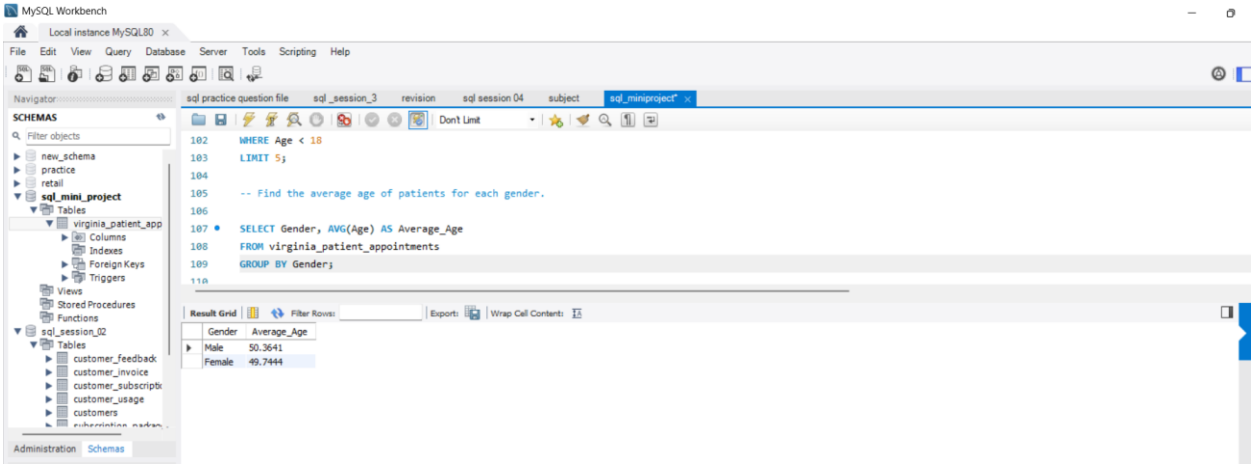
Summary / Interpretation

The 5 youngest patients in the dataset are all under 18 years old, with the following ages: 2, 5, 6, 10, and 14. Among these, 3 are female and 2 are male. The youngest patient, aged 2, is a female.

Aggregation & Grouping

11. Find the average age of patients for each gender.

```
SELECT Gender, AVG(Age) AS Average_Age
FROM virginia_patient_appointments
GROUP BY Gender;
```

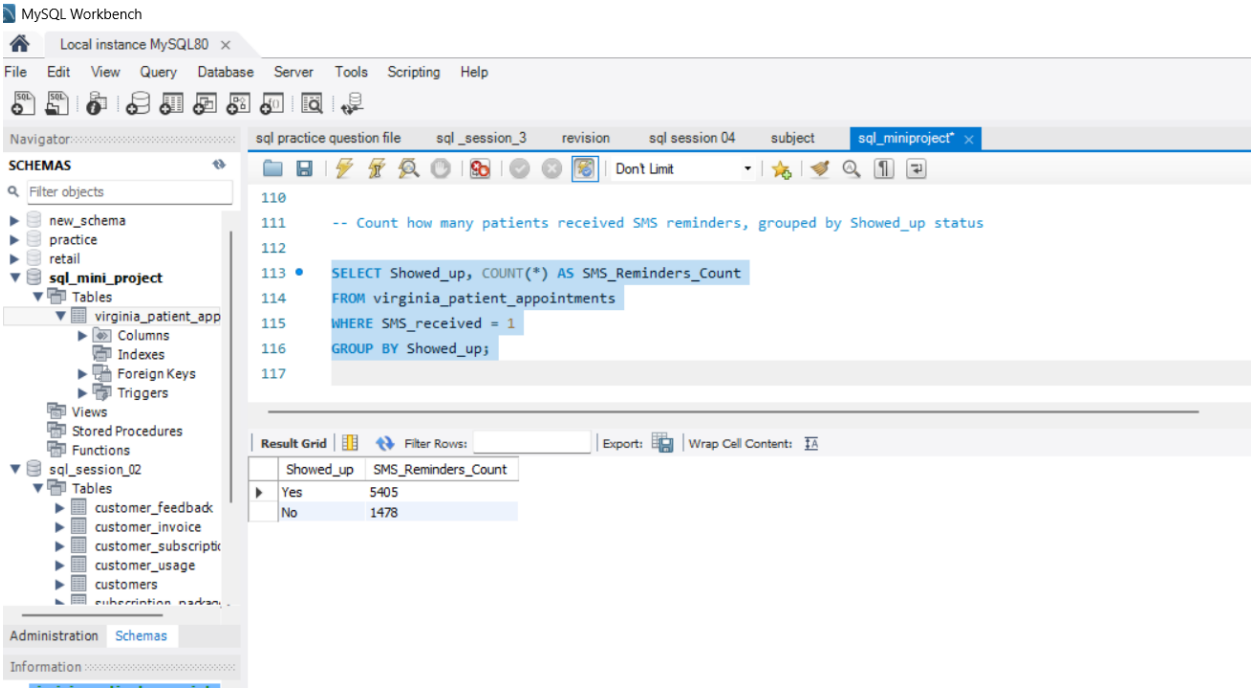


Summary / Interpretation

The average age of patients in the clinic is slightly higher for males (50.36 years) compared to females (49.74 years). This indicates a small age difference between male and female patients, which could be useful for tailoring healthcare services and resource allocation.

12. Count how many patients received SMS reminders, grouped by Showed_up status

```
SELECT Showed_up, COUNT (*) AS SMS_Reminders_Count
FROM virginia_patient_appointments
WHERE SMS_received = 1
GROUP BY Showed_up;
```



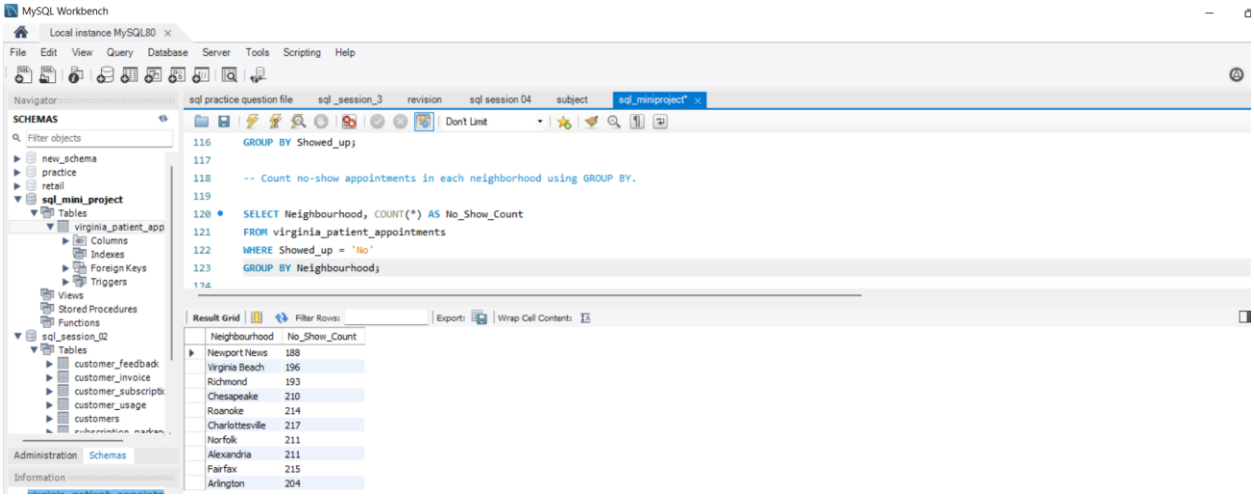
Summary / Interpretation

A total of 5405 patients who received SMS reminders showed up for their appointments, while 1478 patients who received SMS reminders did not show up.

This suggests that SMS reminders may have a positive impact on patient attendance, with a higher number of patients attending their appointments after receiving reminders.

13. Count no-show appointments in each neighborhood using GROUP BY.

```
SELECT Neighbourhood, COUNT (*) AS No_Show_Count
FROM virginia_patient_appointments
WHERE showed_up = 'No'
GROUP BY Neighbourhood;
```

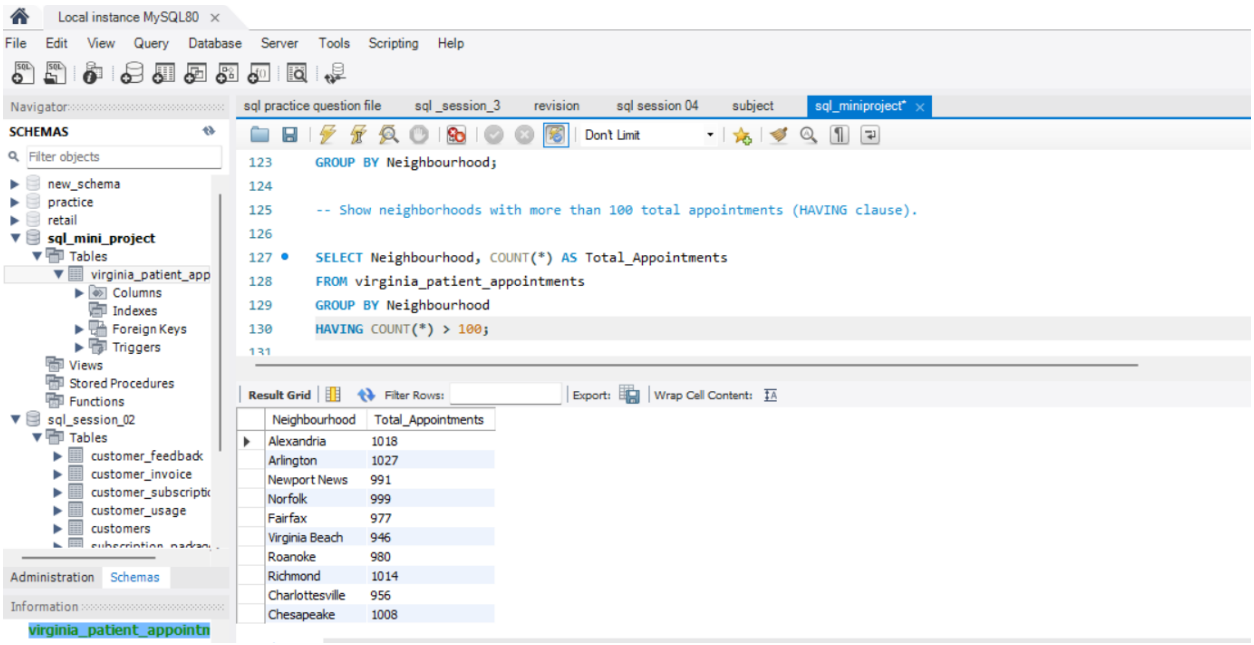


Summary / Interpretation

The no-show appointments vary across neighborhoods, with Charlottesville having the highest count (217) and Newport News the lowest (188). Other neighborhoods like Roanoke (214) and Fairfax (215) also report high no-show rates.

14. **Show neighborhoods with more than 100 total appointments (HAVING clause).**

```
SELECT Neighbourhood, COUNT (*) AS Total Appointments
FROM virginia_patient_appointments
GROUP BY Neighbourhood
HAVING COUNT (*) > 100;
```

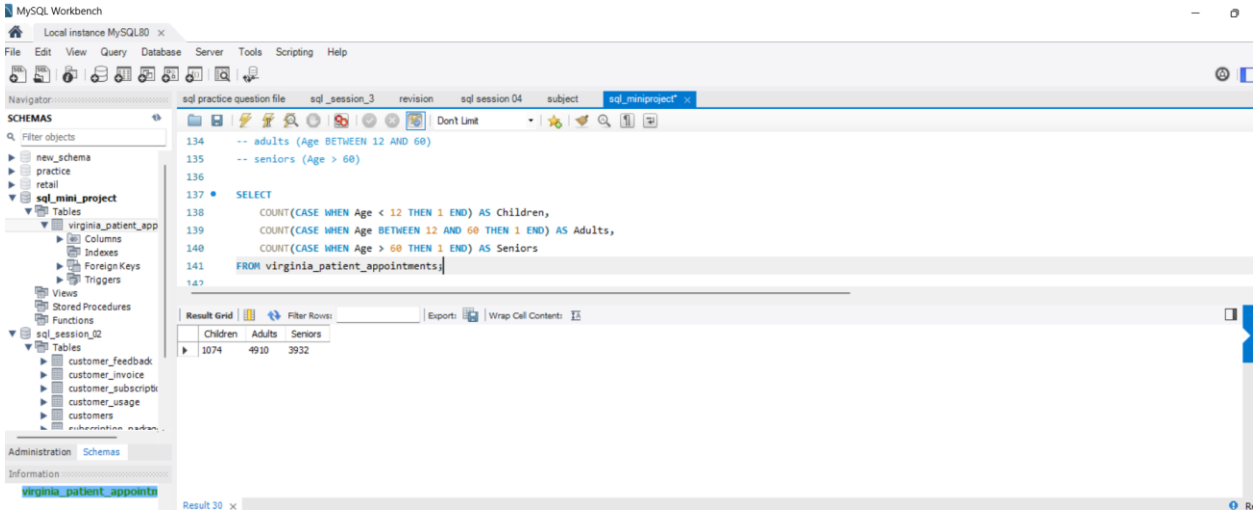


Summary/interpretation

Neighborhoods with more than 100 appointments include Alexandria (1018), Arlington (1027), and Richmond (1014), among others. These areas have higher patient activity, suggesting increased demand for healthcare services.

15. Use CASE to calculate the total number of: children (Age < 12), adults (Age BETWEEN 12 AND 60), seniors (Age > 60)

```
SELECT
COUNT (CASE WHEN Age < 12 THEN 1 END) AS Children,
COUNT (CASE WHEN Age BETWEEN 12 AND 60 THEN 1 END) AS Adults,
COUNT (CASE WHEN Age > 60 THEN 1 END) AS Seniors
FROM virginia_patient_appointments;
```



Summary/ Interpretation

The total number of patients in the dataset is broken down as follows:

- **Children (Age < 12)**: **1,074** patients
- **Adults (Age 12 to 60)**: **4,910** patients
- **Seniors (Age > 60)**: **3,932** patients

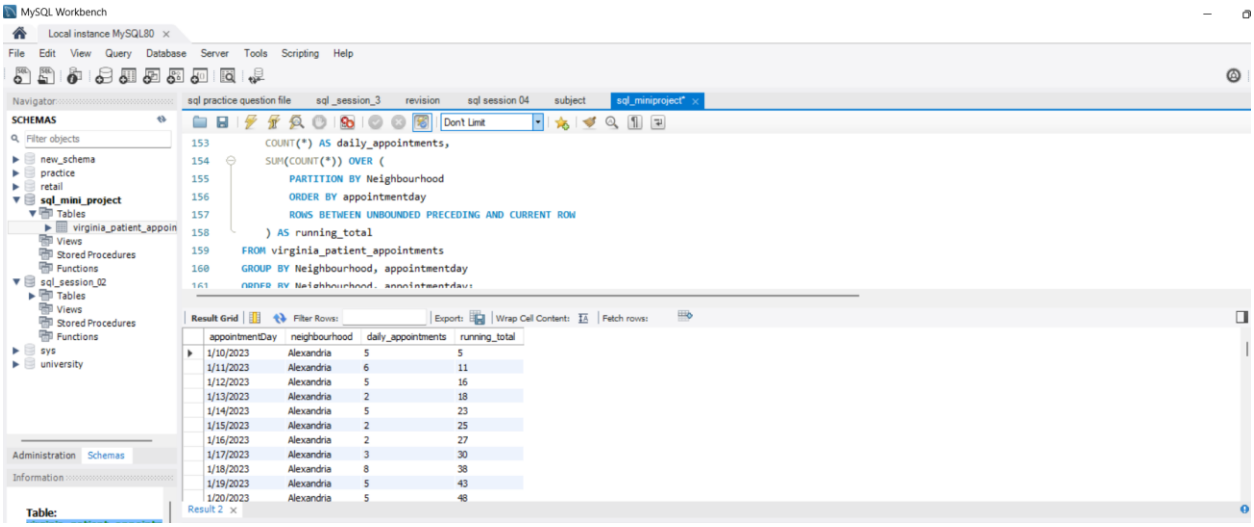
This shows a relatively high proportion of **adults**, followed by **seniors**, with **children** making up a smaller segment of the patient population.

Window Functions

16. **Tracks how appointments accumulate over time in each neighborhood. (Running Total of Appointments per Day) In simple words: How many appointments were there each day and how do the total appointments keep adding up over time in each neighborhood?**

```
SELECT
appointmentDay,
neighbourhood,
COUNT (*) AS daily_appointments,
```

```
SUM (COUNT (*)) OVER (
PARTITION BY Neighbourhood
ORDER BY appointmentday
ROWS BETWEEN UNBOUNDED PRECEDING AND CURRENT ROW
) AS running_total
FROM virginia_patient_appointments
GROUP BY Neighbourhood, appointmentday
ORDER BY Neighbourhood, appointmentday;
```



Summary of Query Output

- **Total Rows Returned:** 1,890

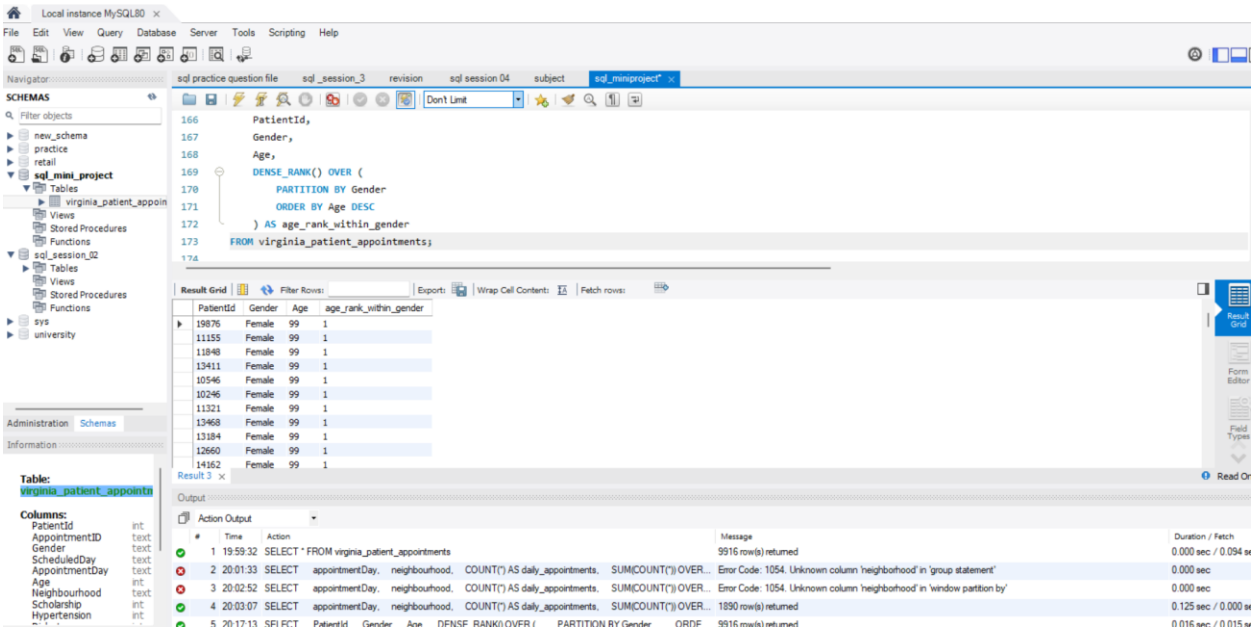
This means there are 1,890 unique combinations of appointment dates and neighborhoods across the 10,000 appointments in dataset.

- **Columns Included:**

1. **appointmentDay** – The date on which the appointment was scheduled to happen.
2. **neighbourhood** – The area where the patient resides.
3. **daily_appointments** – Number of appointments held in that neighbourhood on that day.
4. **running_total** – Cumulative total of appointments in that neighbourhood up to and including that day.

17. Use Dense_Rank () to rank patients by age within each gender group.

```
SELECT
PatientId,
Gender,
Age,
DENSE_RANK () OVER (
PARTITION BY Gender
ORDER BY Age DESC
) AS age_rank_within_gender
FROM virginia_patient_appointments;
```

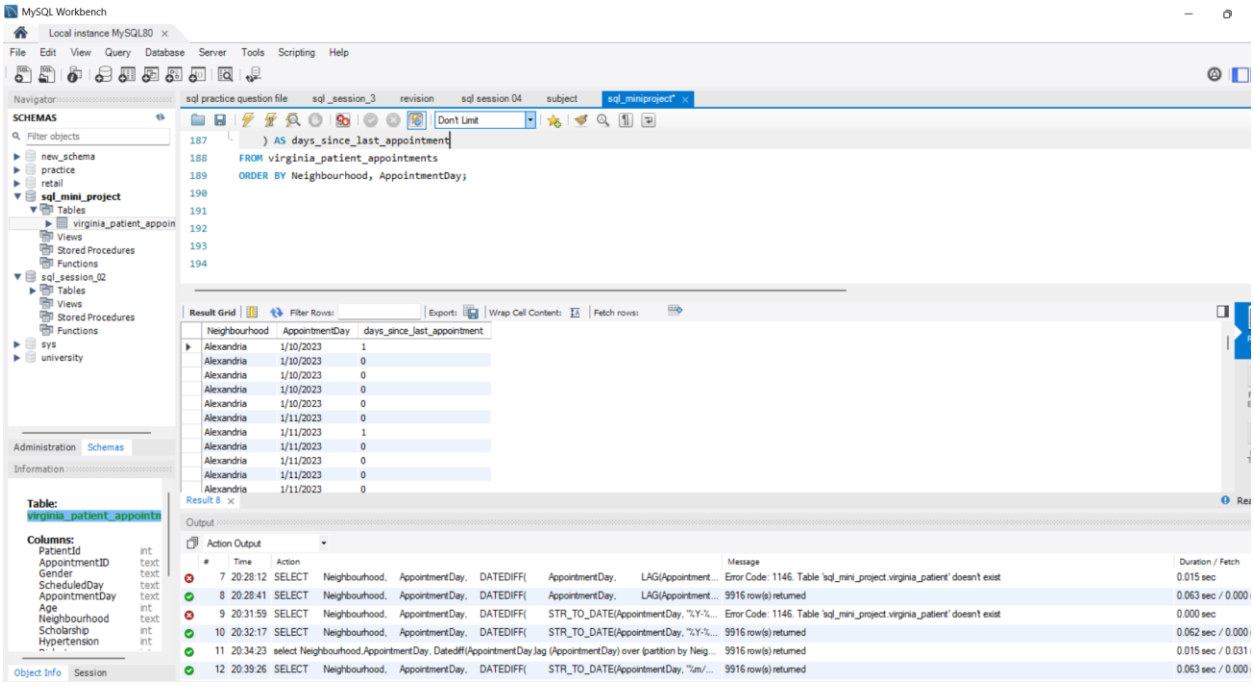


Summary and Interpretation

The query ranked 9,916 patients by age within each gender group using the DENSE_RANK () function, with the oldest patients receiving the highest rank in each gender category. This ranking helps reveal age distribution patterns among male and female patients, highlighting which age groups are most common within each gender. Since patients with the same age receive the same rank without gaps, it allows for accurate grouping and comparison.

18. How many days have passed since the last appointment in the same neighborhood? (Hint: DATEDIFF and Lag) (This helps to see how frequently appointments are happening in each neighborhood.)

```
SELECT
Neighbourhood,
AppointmentDay,
DATEDIFF (
STR_TO_DATE (AppointmentDay, '%m/%d/%y'),
LAG (STR_TO_DATE (AppointmentDay, '%m/%d/%y')) OVER (
PARTITION BY Neighbourhood
ORDER BY STR_TO_DATE (AppointmentDay, '%m/%d/%y')
)) AS days_since_last_appointment
FROM virginia_patient_appointments
ORDER BY Neighbourhood, AppointmentDay;
```



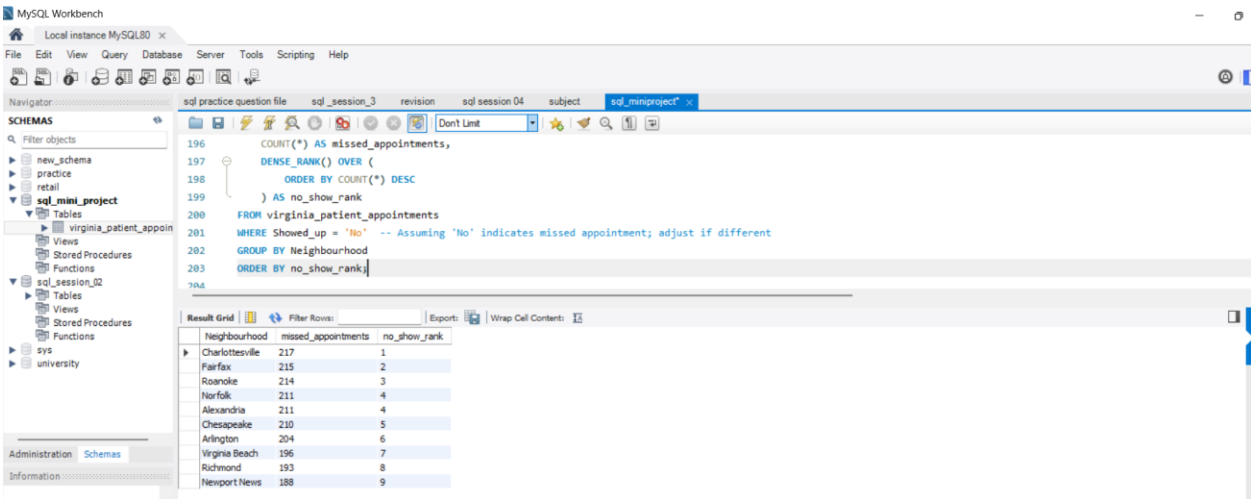
Summary and Interpretation

The query returned 9,916 rows showing the number of days elapsed since the previous appointment in the same neighbourhood. The values range from **0, 1, and higher**, where:

- **0 days** means the current appointment is on the **same day** as the previous appointment in that neighbourhood.
- **1 day** means the appointment is exactly **one day after** the previous one.
- Higher numbers indicate longer gaps between appointments.

19. Which neighborhoods have the highest number of missed appointments? Use DENSE_RANK () to rank neighborhoods based on the number of no-show appointments.

```
SELECT
Neighbourhood,
COUNT (*) AS missed_appointments,
DENSE_RANK () OVER (
ORDER BY COUNT (*) DESC
) AS no_show_rank
FROM virginia_patient_appointments
WHERE showed_up = 'No' -- Assuming 'No' indicates missed appointment; adjust if different
GROUP BY Neighbourhood
ORDER BY no_show_rank;
```



Summary and Interpretation

The analysis ranks neighborhoods in Virginia based on the number of missed medical appointments. Charlottesville tops the list with 217 no-shows, indicating it has the highest rate of patients missing their scheduled visits. Fairfax and Roanoke follow closely with 215 and 214 missed appointments, respectively. Some neighborhoods, like Norfolk and Alexandria, share the same number of no-shows (211), resulting in a tied rank.

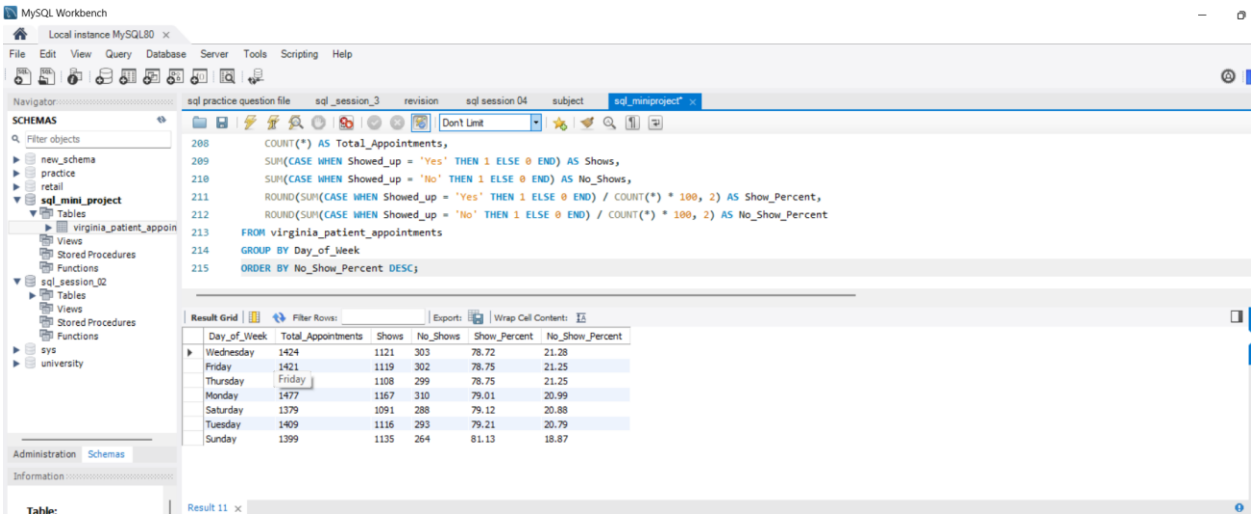
Complex Query

20. Are patients more likely to miss appointments on certain days of the week?

Steps to follow for question # 20

- **(Use the AppointmentDay column in function dayname () to extract the day name (like Monday, Tuesday, etc.).**
- **Count how many appointments were scheduled, how many showed up (showed_up = "yes") and how many were missed (Showed_up = 'No') on each day.**
- **Calculate the percentage of shows and no-shows for better comparison between days.**
- **Formula: (count of Showed_up = 'yes' / total appointment count) * 100, Use round function to show upto two decimal points**
- **Sort the result by No_Show_Percent in descending order to see the worst-performing days first.**

```
SELECT
    DAYNAME (STR_TO_DATE (AppointmentDay, '%m/%d/%y')) AS Day_of_Week,
    COUNT (*) AS Total_Appointments,
    SUM (CASE WHEN Showed_up = 'Yes' THEN 1 ELSE 0 END) AS Shows,
    SUM (CASE WHEN Showed_up = 'No' THEN 1 ELSE 0 END) AS No_Shows,
    ROUND (SUM (CASE WHEN Showed_up = 'Yes' THEN 1 ELSE 0 END) / COUNT (*) * 100, 2) AS Show_Percent,
    ROUND (SUM (CASE WHEN Showed_up = 'No' THEN 1 ELSE 0 END) / COUNT (*) * 100, 2) AS No_Show_Percent
FROM virginia_patient_appointments
GROUP BY Day_of_Week
ORDER BY No_Show_Percent DESC;
```



Day_of_Week	Total_Appointments	Shows	No_Shows	Show_Percent	No_Show_Percent
Wednesday	1424	1121	303	78.72	21.28
Friday	1421	1119	302	78.75	21.25
Thursday	1408	1108	299	78.75	21.25
Monday	1477	1167	310	79.01	20.99
Saturday	1379	1091	288	79.12	20.88
Tuesday	1409	1116	293	79.21	20.79
Sunday	1399	1135	264	81.13	18.87

Summary and Interpretation

The analysis of appointment attendance by day of the week reveals that **Wednesday, Friday, and Thursday** have the **highest no-show rates**, each with over **21%** of patients missing their appointments. Specifically, **Wednesday** has the highest no-show percentage at **21.28%**, followed closely by Friday and Thursday. In contrast, **Sunday** has the **lowest no-show rate at 18.87%**, indicating better patient attendance. These patterns suggest that patients may be more likely to skip appointments on weekdays, especially midweek, possibly due to work or personal commitments.

4. Key Findings

Age Group Distribution

- Children (<12 years): 1,074 patients
- Adults (12–60 years): 4,910 patients
- Seniors (>60 years): 3,932 patients
  Adults make up the largest portion, followed by seniors, indicating a middle-to-older age-skewed patient population.

Neighborhoods with High Appointment Volume

- Neighborhoods like **Arlington (1,027)**, **Alexandria (1,018)**, and **Richmond (1,014)** recorded over 1,000 appointments, indicating high demand for healthcare services in these regions.

No-Show Analysis by Neighborhood

- **Charlottesville (217)**, **Fairfax (215)**, and **Roanoke (214)** had the highest number of missed appointments, signaling a need for targeted intervention in these areas.

Impact of SMS Reminders

- 5,405 patients with SMS reminders **attended**, while 1,478 **did not** — suggesting SMS reminders help improve attendance rates.

Gender-Based Age Insights

- Average age: **Males – 50.36 years**, **Females – 49.74 years**
  Males are slightly older on average, indicating potential gender-based differences in health conditions or appointment behavior.

Youngest and Oldest Patients

- 5 youngest patients: Ages **2, 5, 6, 10, 14**
- 5 oldest patients: All aged **99**
  Elderly patients (especially those 99 years old) form a notable segment requiring age-specific care.

Chronic Conditions (Diabetes & Hypertension)

- 435 patients had **both conditions**, highlighting a significant high-risk group for focused healthcare management.

Appointment Attendance Labeling

- New column AppointmentStatus was added for **9,916** records, showing most patients **attended** their appointments.

Data Accuracy Improvements

- Showed_up status was corrected for **4 null/missing** entries, improving dataset completeness.

Appointment Trend in May 2023

- **1,717** appointments were recorded in May 2023, giving a clear time-bound snapshot of clinic activity.

Female Patients and SMS

- **3,465** female patients received SMS reminders — showing proactive communication efforts toward women.

Geographic Spread of Patients

- Patients came from various cities, including **Richmond**, **Norfolk**, **Virginia Beach**, and **Charlottesville**, reflecting the clinic's wide regional reach.

Elderly Patients and Appointment Adherence

- First 10 appointments for patients over 60 showed 100% attendance, indicating strong adherence among senior patients.

Running Total of Appointments

- A cumulative count of appointments over time by neighborhood reveals trends in healthcare demand and helps in staffing/resource planning.

Age Ranking by Gender

- DENSE_RANK () highlighted the oldest patients within each gender, aiding in age-specific program planning.

Appointment Frequency in Neighborhoods

- Most neighborhoods had 1–day gaps between appointments, indicating high consistency and frequency of healthcare access.

Missed Appointments by Day of Week

- Highest no-show rates: **Wednesday (21.28%)**, **Friday**, and **Thursday**
- Lowest no-show: **Sunday (18.87%)** — suggests weekdays, especially midweek, are more prone to missed appointments.

Better Attendance on Weekends

- Patients are slightly more likely to attend on **Saturday and Sunday**, offering scheduling optimization opportunities.

5. Recommendations

Based on the insights gathered from the data analysis, the following targeted strategies are recommended to help **reduce missed appointments** and improve overall clinic efficiency:

1. **Enhance SMS Reminder Campaigns**
    - Since a **higher attendance rate** was observed among patients who received SMS reminders, ensure **automated and timely SMS notifications** are sent before every appointment.
    - Consider **personalized reminders** mentioning date, time, and consequences of missing the appointment.
2. **Optimize Scheduling on High No-Show Days**
    - **Wednesday, Thursday, and Friday** had the **highest no-show rates**. Reduce scheduling on these days for first-time or high-risk no-show patients.
    - Alternatively, **double-confirm appointments** on these days or overbook slightly to balance expected absences.
3. **Encourage Weekend Appointments for Certain Patients**
    - Show-up rates are higher on **Saturday and Sunday**. Offer weekend slots to patients with a history of no-shows or working adults who may find weekday appointments difficult.
4. **Target High No-Show Neighborhoods**
    - Implement outreach or follow-up calls in **Charlottesville, Fairfax, Roanoke, and Norfolk**, where no-shows were most common.
    - Introduce **incentives or loyalty programs** for consistent attendance in these areas.

5. **Provide Extra Support for High-Risk Groups**
    - Elderly patients and those with **both diabetes and hypertension** may benefit from transportation assistance, family notifications, or caregiver alerts to ensure attendance.
6. **Gender-Specific Communication**
    - Slight age differences by gender suggest different health engagement levels. Customize **reminders and messaging** styles to suit male and female patients more effectively.
7. **Introduce Appointment Confirmation System**
    - Use a **"Confirm or Reschedule" SMS link** or phone call follow-up to allow patients to confirm attendance, reducing the likelihood of unattended slots.

6. Conclusion

The analysis of 10,000 medical appointment records from Virginia reveals key patterns in patient attendance. Missed appointments are influenced by factors such as day of the week, neighborhood, and SMS reminders. Most patients who received reminders attended their appointments, and attendance was generally higher on weekends. Certain neighborhoods like Charlottesville and Fairfax showed higher no-show rates, while older patients and those with chronic conditions displayed relatively better adherence. These insights can guide healthcare providers in improving appointment scheduling, communication strategies, and patient engagement to reduce no-show rates and enhance service delivery.