**SQL Data Analysis Project Report**

Submitted as part of Internship at SoulVibe Tech

**Submitted By:  
Ajay Kumar Sahu**

Date: 25 June, 2025

# Data Cleaning and Setup

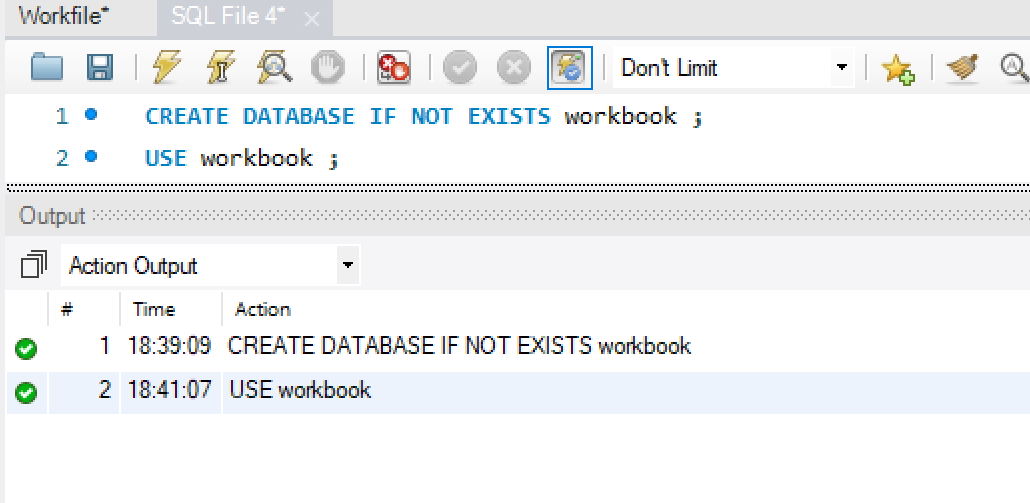
**Firstly , I will show how I have cleaned the data and imported the .csv file**

* I have used the Power Query Editor to replace the empty cells to NA(Not Available)
* Then I have changed some column names i.e. (sr\_no, district, taluka, college\_name, university, college\_type, course\_name, course\_type, is\_professional, course, course\_duration, course\_category)
* Then I have Saved the file named as Workfile.csv in the 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/Workfile.csv' So that there is no error or restriction in importing the CSV file to MySQL.
* Then I have Created the DATABASE named workbook by the Query –

**Query -**

CREATE DATABASE IF NOT EXISTS workbook ;

USE workbook ;



* Then, I have created the table (table name – college\_courses) by –

**Query –**

CREATE TABLE college\_courses (

sr\_no INT,

district VARCHAR(100),

taluka VARCHAR(100),

college\_name VARCHAR(255),

university VARCHAR(255),

college\_type VARCHAR(100),

course\_name TEXT,

course\_type VARCHAR(100),

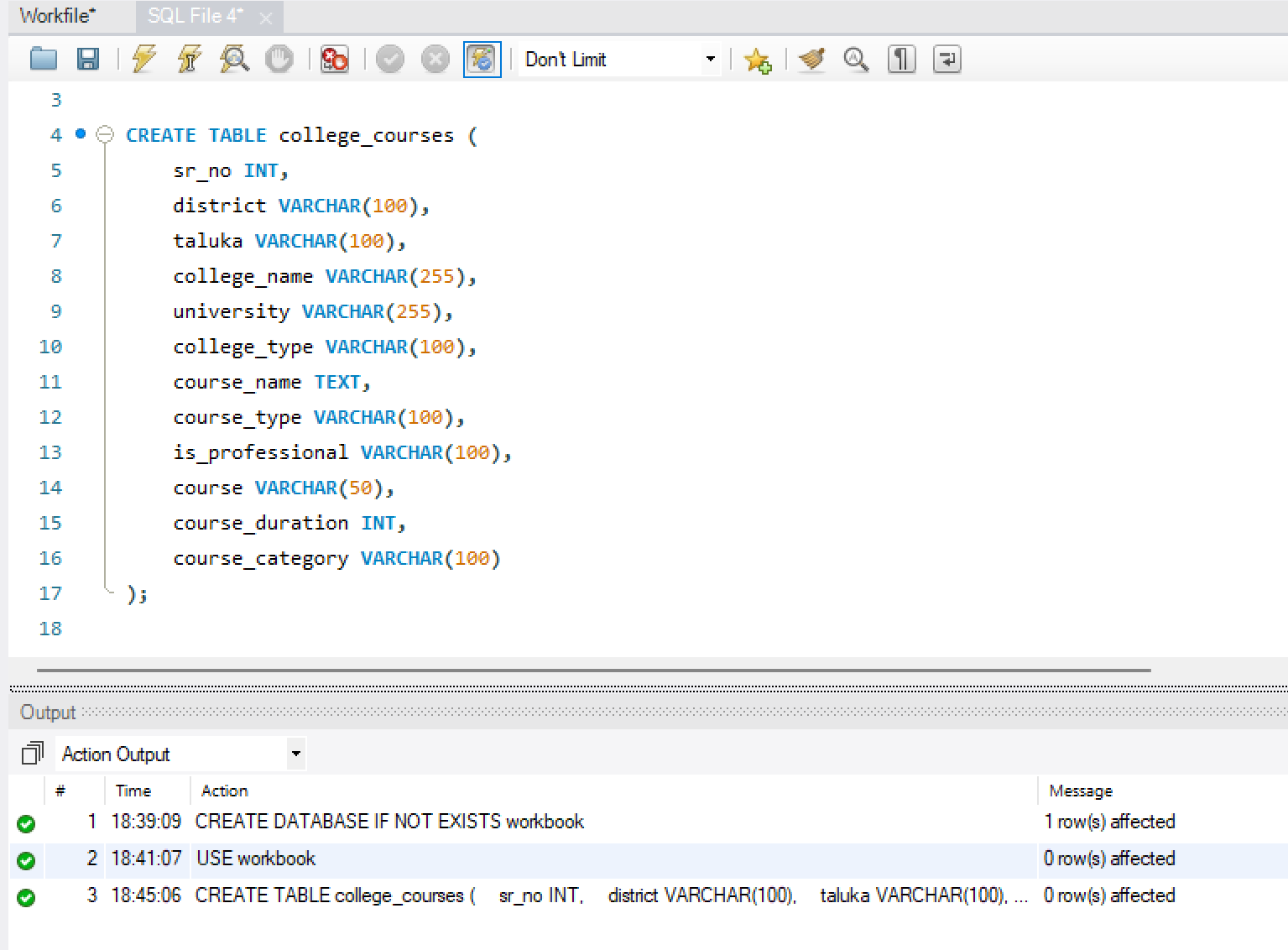
is\_professional VARCHAR(100),

course VARCHAR(50),

course\_duration INT,

course\_category VARCHAR(100)

);



* Then I have imported the Workbook.csv file in MySQL by –

**Query –**

1. SET GLOBAL local\_infile = 1;

2. LOAD DATA INFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/Workfile.csv'

INTO TABLE college\_courses

FIELDS TERMINATED BY ','

ENCLOSED BY '"'

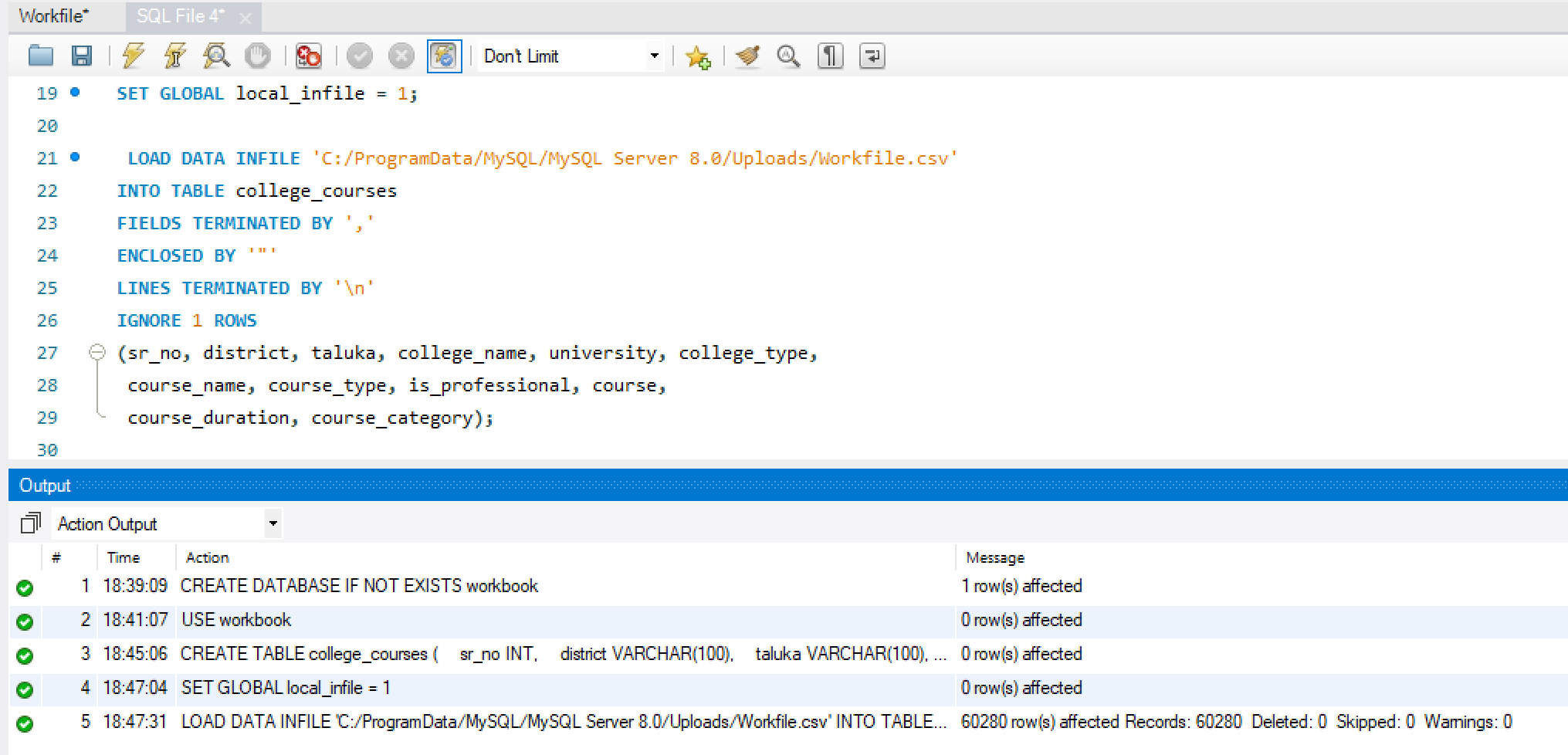
LINES TERMINATED BY '\n'

IGNORE 1 ROWS

(sr\_no, district, taluka, college\_name, university, college\_type,

course\_name, course\_type, is\_professional, course,

course\_duration, course\_category);

****

* I have also remove the row limit by going to Edit 🡪 Preferences 🡪 SQL Execution 🡪 Uncheck the Limit Rows.

SQL Questions

Q1. Find the top 5 districts with the highest number of colleges offering professional courses.

**Query –**

-- Count how many colleges offer professional courses in each district

**SELECT**

district,

**COUNT**(**DISTINCT** college\_name) **AS** total\_professional\_colleges

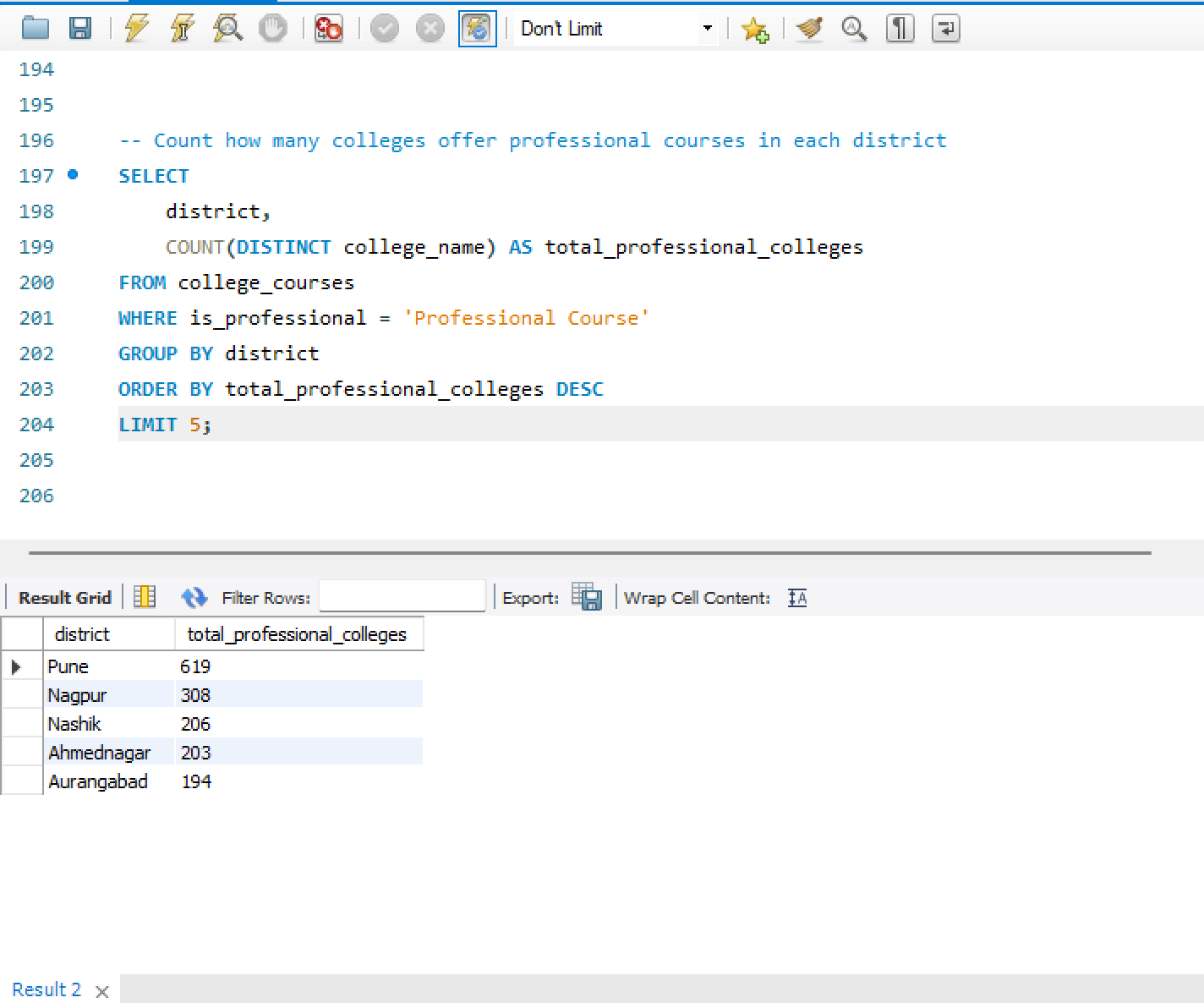
**FROM** college\_courses

**WHERE** is\_professional = 'Professional Course'

**GROUP** **BY** district

**ORDER** **BY** total\_professional\_colleges **DESC**

**LIMIT** 5;



The output shows that districts like Pune, Nagpur, and Nashik have the highest number of professional colleges, indicating these are major educational hubs.

Q2. Calculate the average course duration (in months) for each Course Type and sort them in descending order.

**Query –**

-- Calculate average duration for each course type

**SELECT**

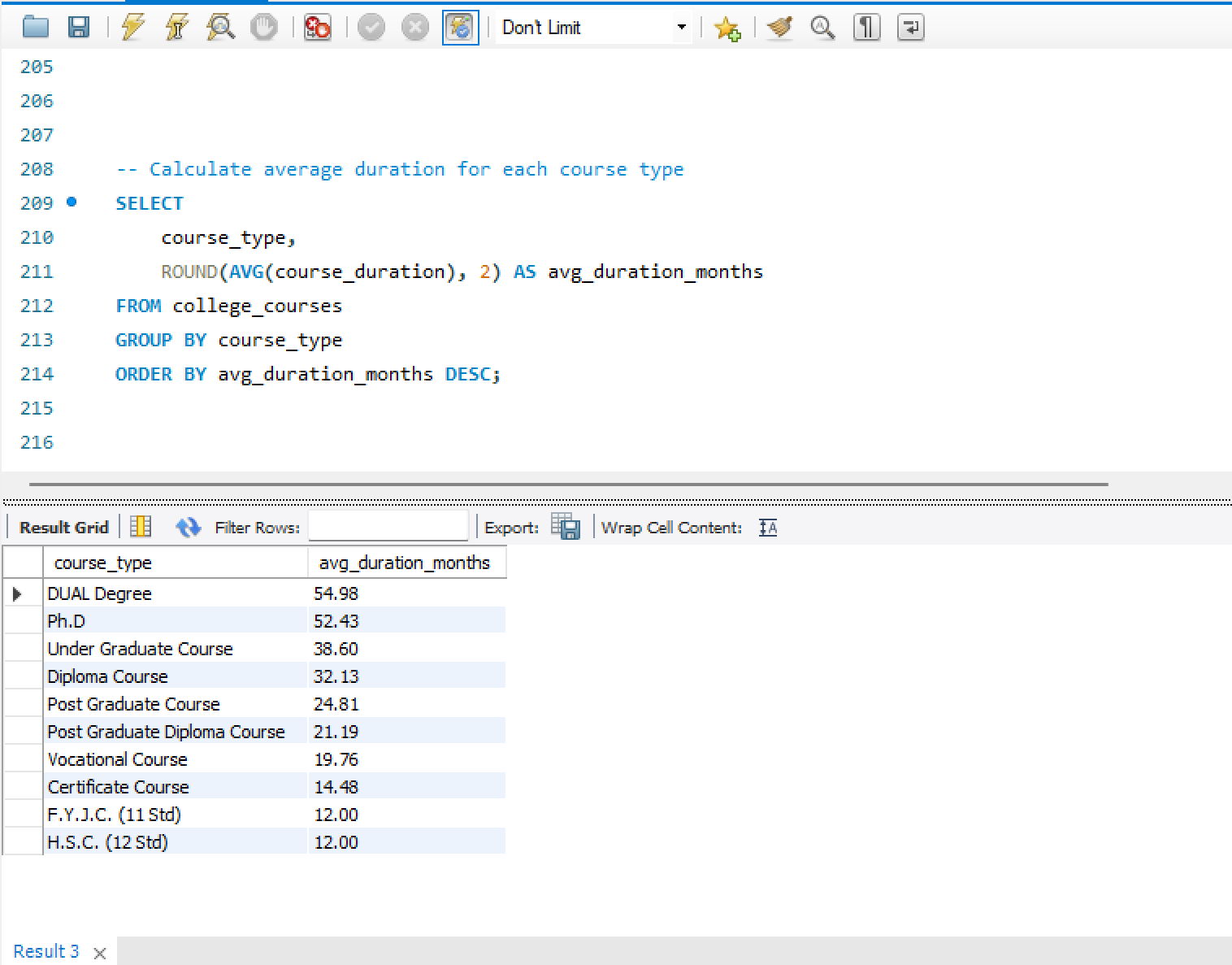
course\_type,

**ROUND**(**AVG**(course\_duration), 2) **AS** avg\_duration\_months

**FROM** college\_courses

**GROUP** **BY** course\_type

**ORDER** **BY** avg\_duration\_months **DESC**;



The average course duration is highest for DUAL Degree and Ph.D programs, while diploma and certificate courses are comparatively shorter, which is expected.

Q3. Count how many unique College Names offer each Course Category.

**Query–**

-- Count unique colleges offering each course category

**SELECT**

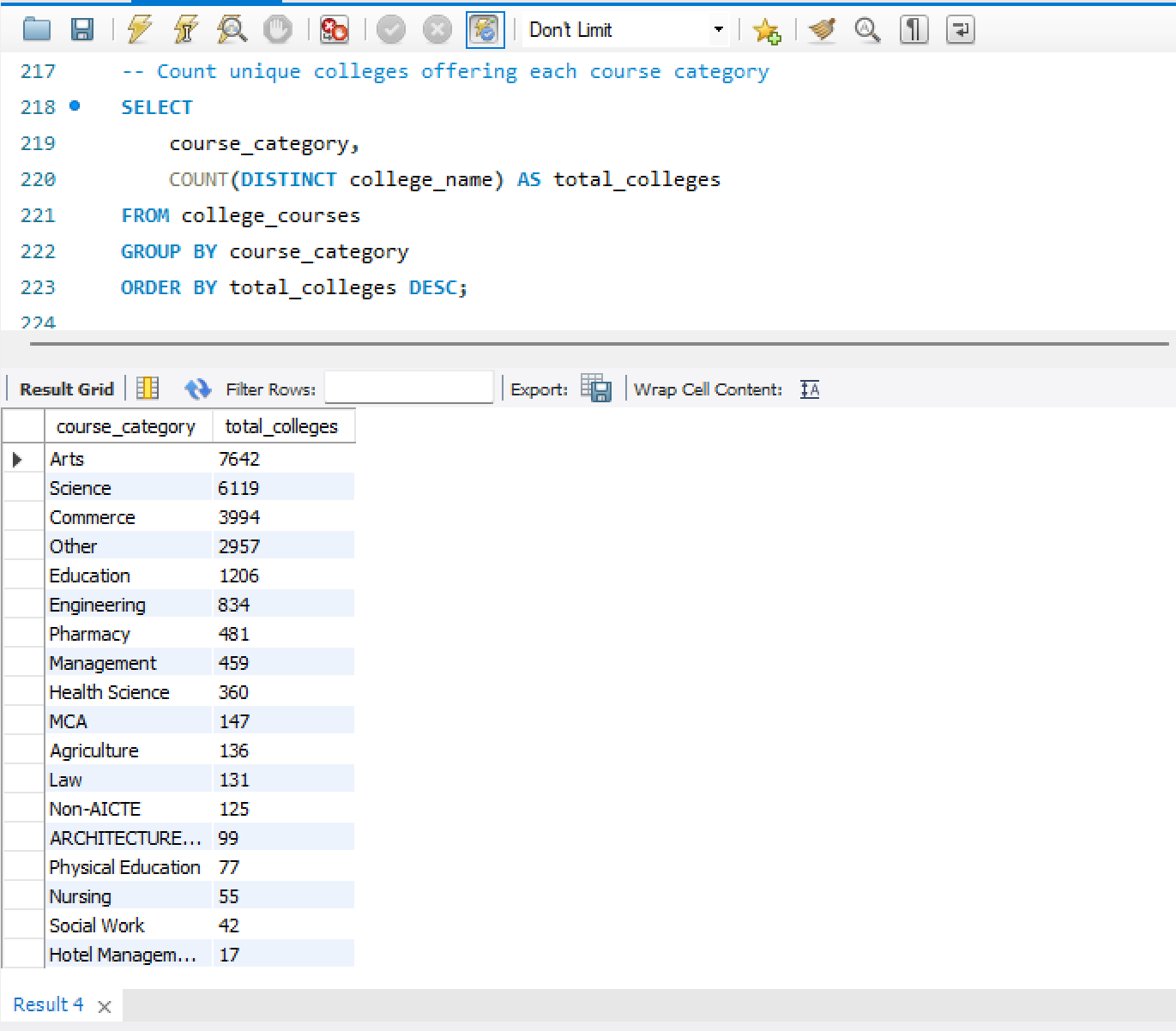
course\_category,

**COUNT**(**DISTINCT** college\_name) **AS** total\_colleges

**FROM** college\_courses

**GROUP** **BY** course\_category

**ORDER** **BY** total\_colleges **DESC**;



Arts and Science categories are offered by the most number of unique colleges, showing their popularity among institutions.

Q4. Find the names of colleges offering both Post Graduate and Under Graduate courses.

**Query –**

-- Get colleges that offer both PG and UG courses

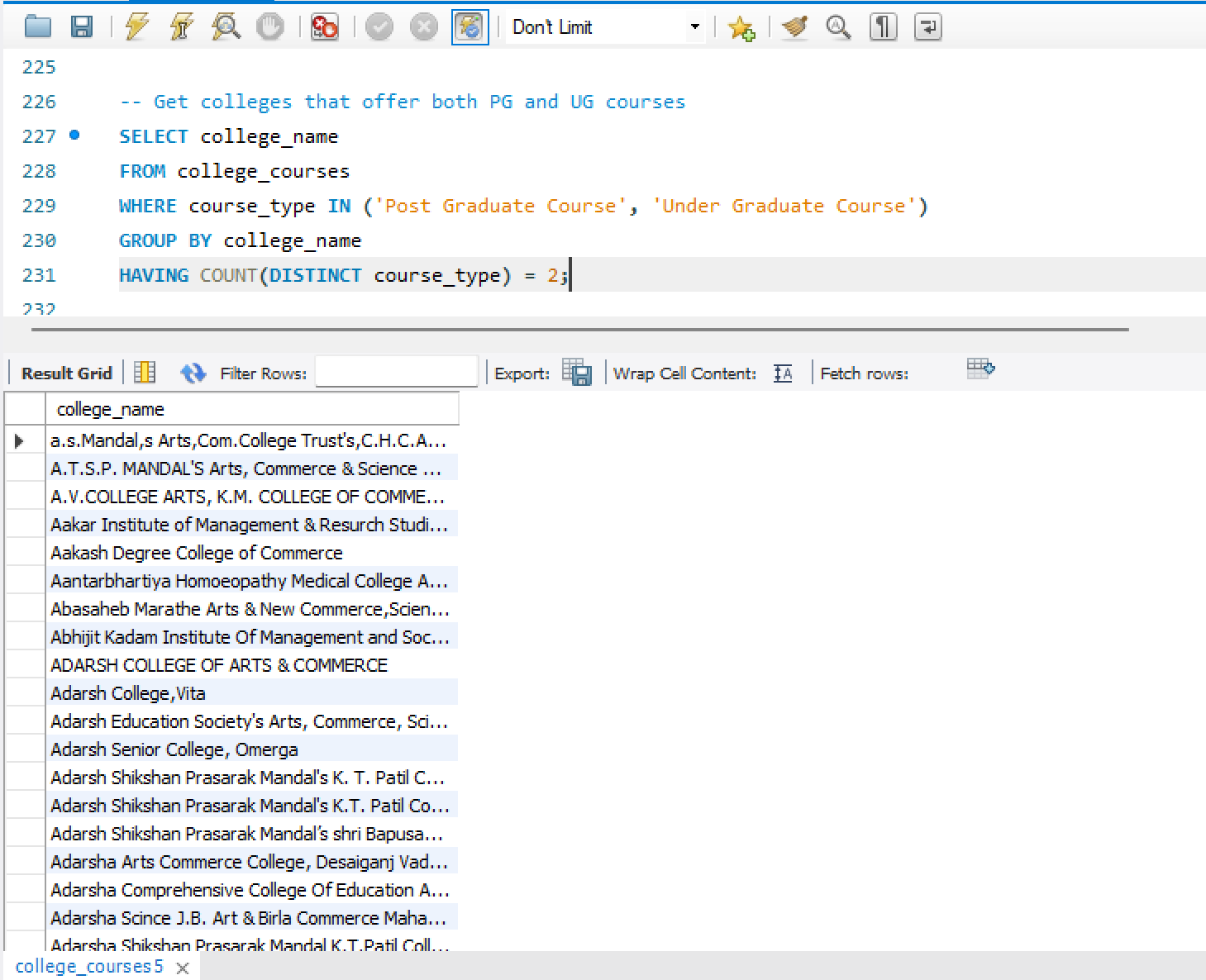
**SELECT** college\_name

**FROM** college\_courses

**WHERE** course\_type **IN** ('Post Graduate Course', 'Under Graduate Course')

**GROUP** **BY** college\_name

**HAVING** **COUNT**(**DISTINCT** course\_type) = 2;



The results highlight multiple colleges that offer both UG and PG courses, which shows that many institutions support advanced academic pathways.

Q5. List all universities that have more than 10 unaided courses that are not professional.

**Query –**

-- Find universities with more than 10 unaided and non-professional courses

**SELECT**

university,

**COUNT**(\*) **AS** total\_courses

**FROM** college\_courses

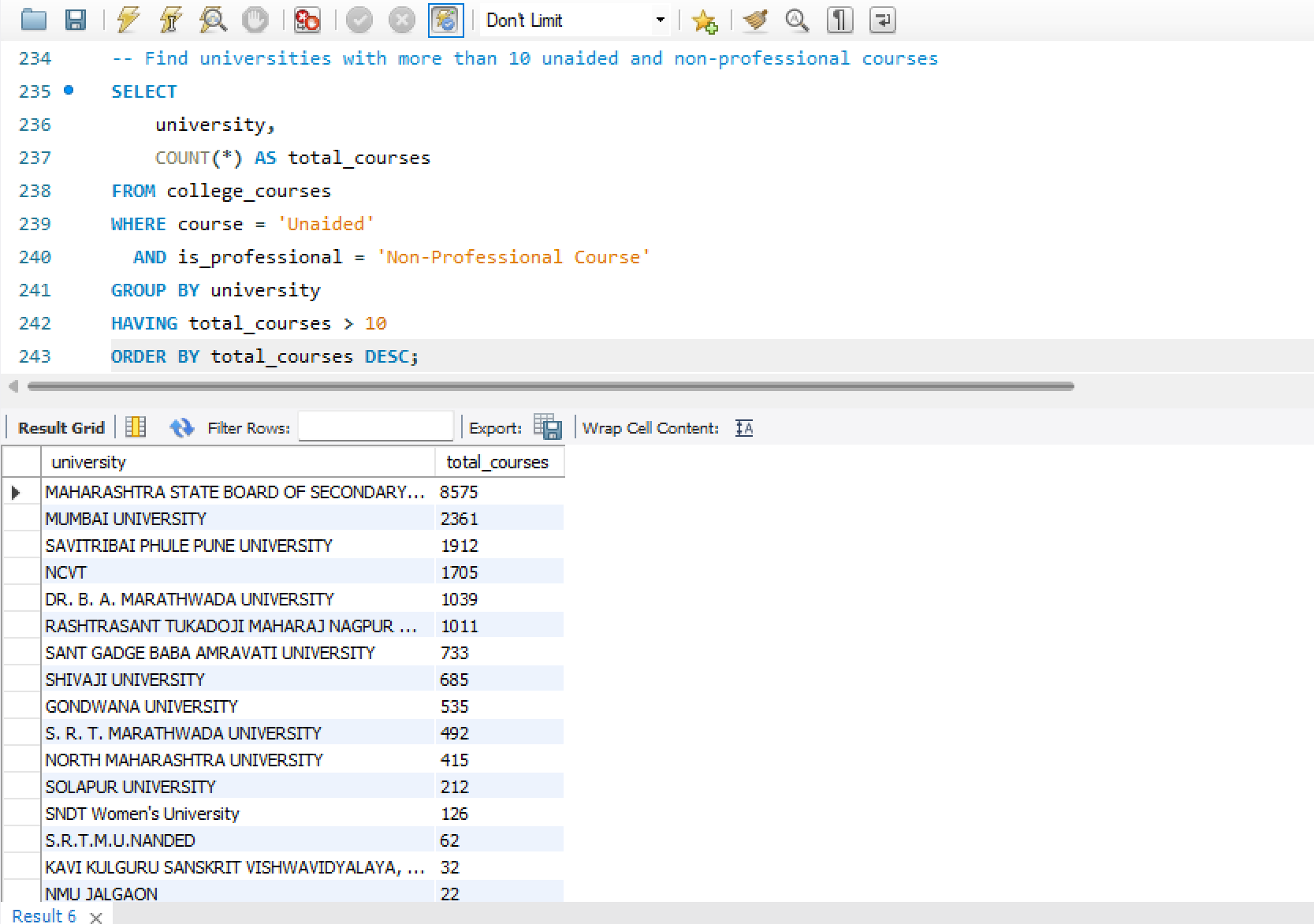
**WHERE** course = 'Unaided'

**AND** is\_professional = 'Non-Professional Course'

**GROUP** **BY** university

**HAVING** total\_courses > 10

**ORDER** **BY** total\_courses **DESC**;



A few universities offer more than 10 unaided, non-professional courses, suggesting their focus on self-financed and general education programs.

Q6. Display colleges from the "Engineering" category that have at least one course with a duration greater than the category’s average.

**Query –**

-- Find colleges in 'Engineering' where at least one course is above average duration

**SELECT** **DISTINCT** college\_name

**FROM** college\_courses

**WHERE** course\_category = 'Engineering'

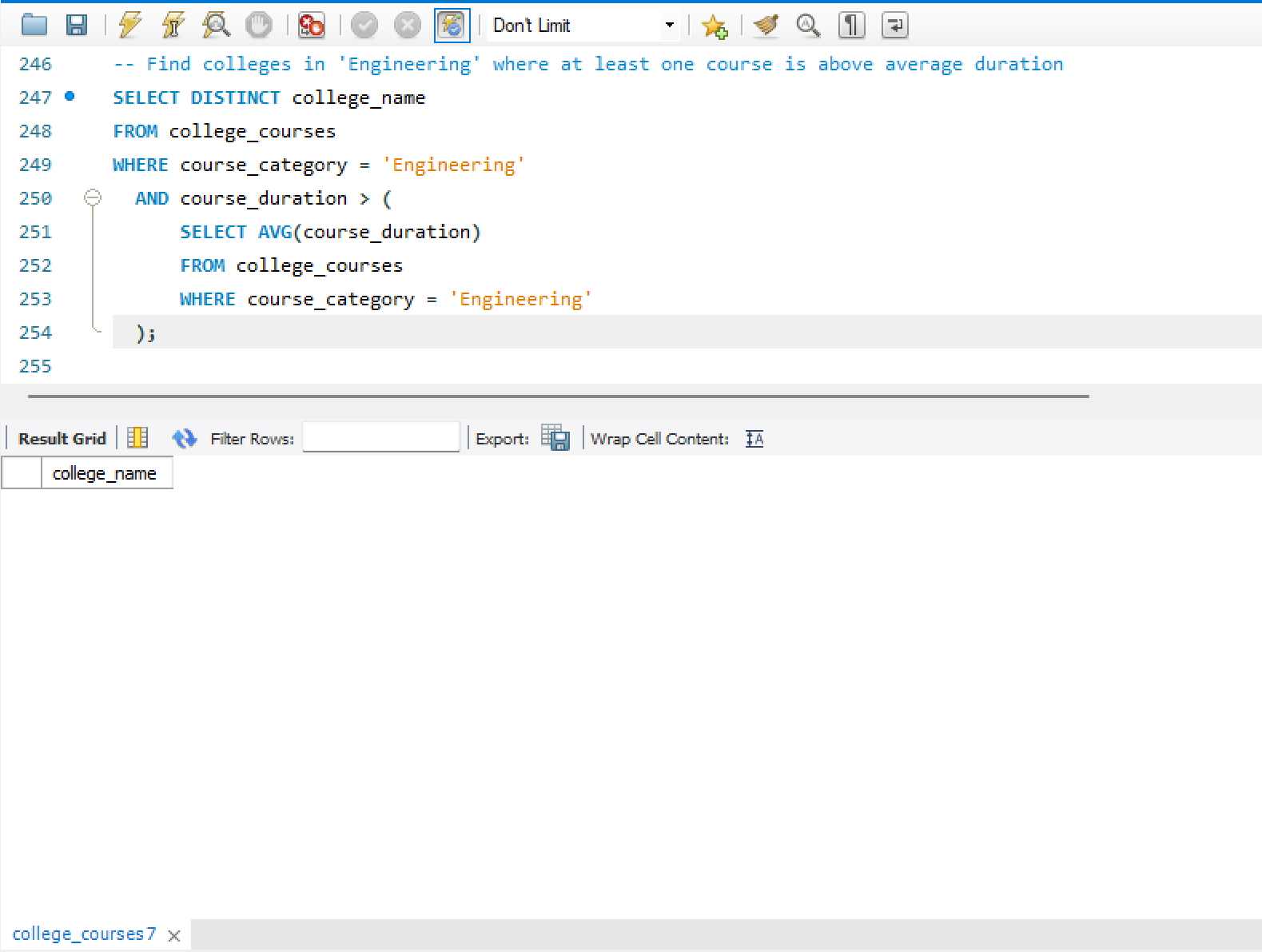
**AND** course\_duration > (

**SELECT** **AVG**(course\_duration)

**FROM** college\_courses

**WHERE** course\_category = 'Engineering'

);



No college in the Engineering category had a course that exceeded the average category duration, meaning most colleges offer similarly structured programs.

Q7. Assign a rank to each course within a College Name based on course duration, longest first.

**Query –**

-- Rank courses within each college based on course duration (longest first)

**SELECT**

college\_name,

course\_name,

course\_duration,

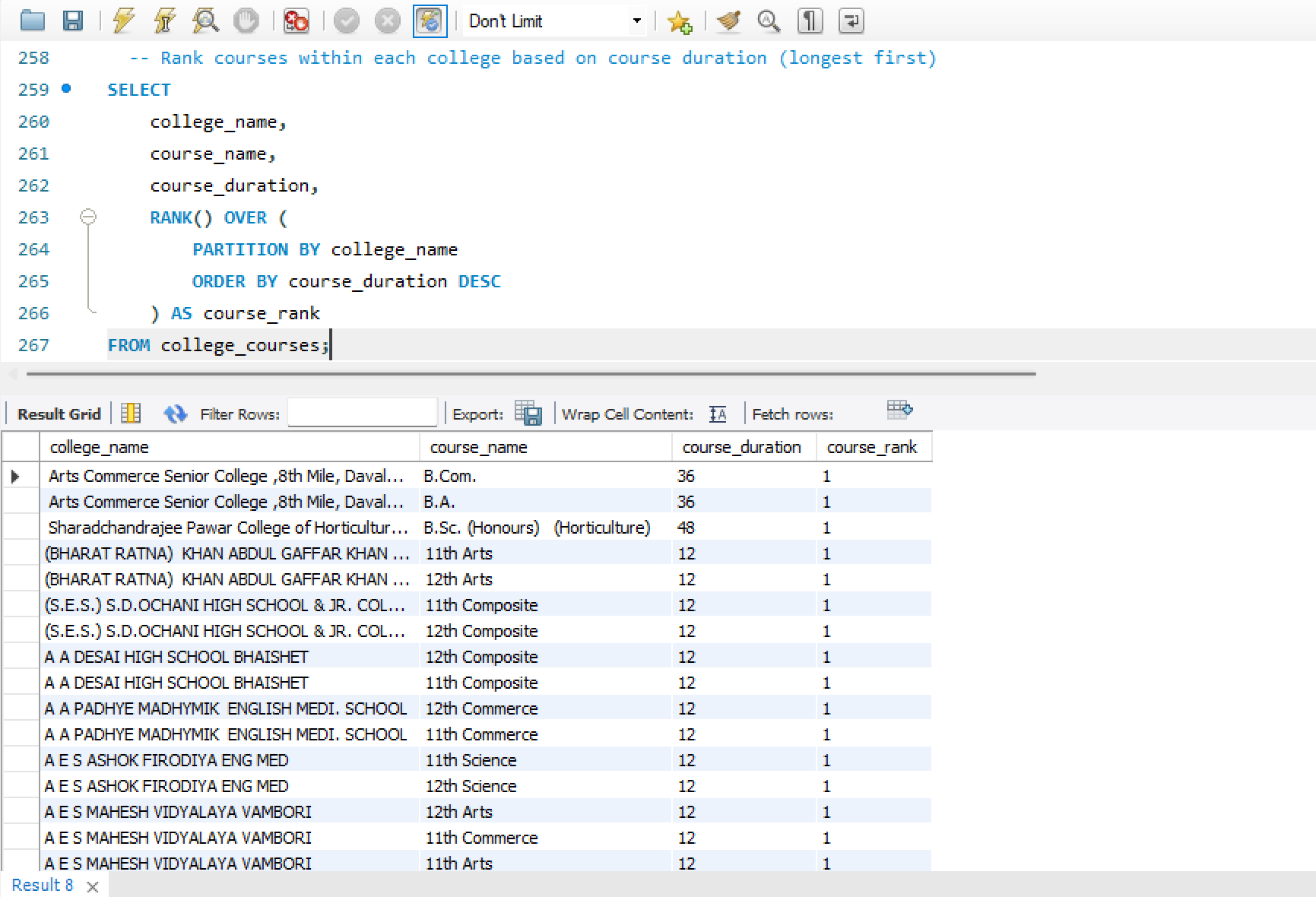
**RANK**() **OVER** (

**PARTITION** **BY** college\_name

**ORDER** **BY** course\_duration DESC

) **AS** course\_rank

**FROM** college\_courses;



The ranking helped show which courses are the longest within each college — typically UG or PG courses rank higher in duration.

Q8. Find colleges where the longest and shortest course durations are more than 24 months apart.

**Query –**

-- Show colleges where course duration gap is more than 24 months

**SELECT**

college\_name,

**MAX**(course\_duration) **AS** max\_duration,

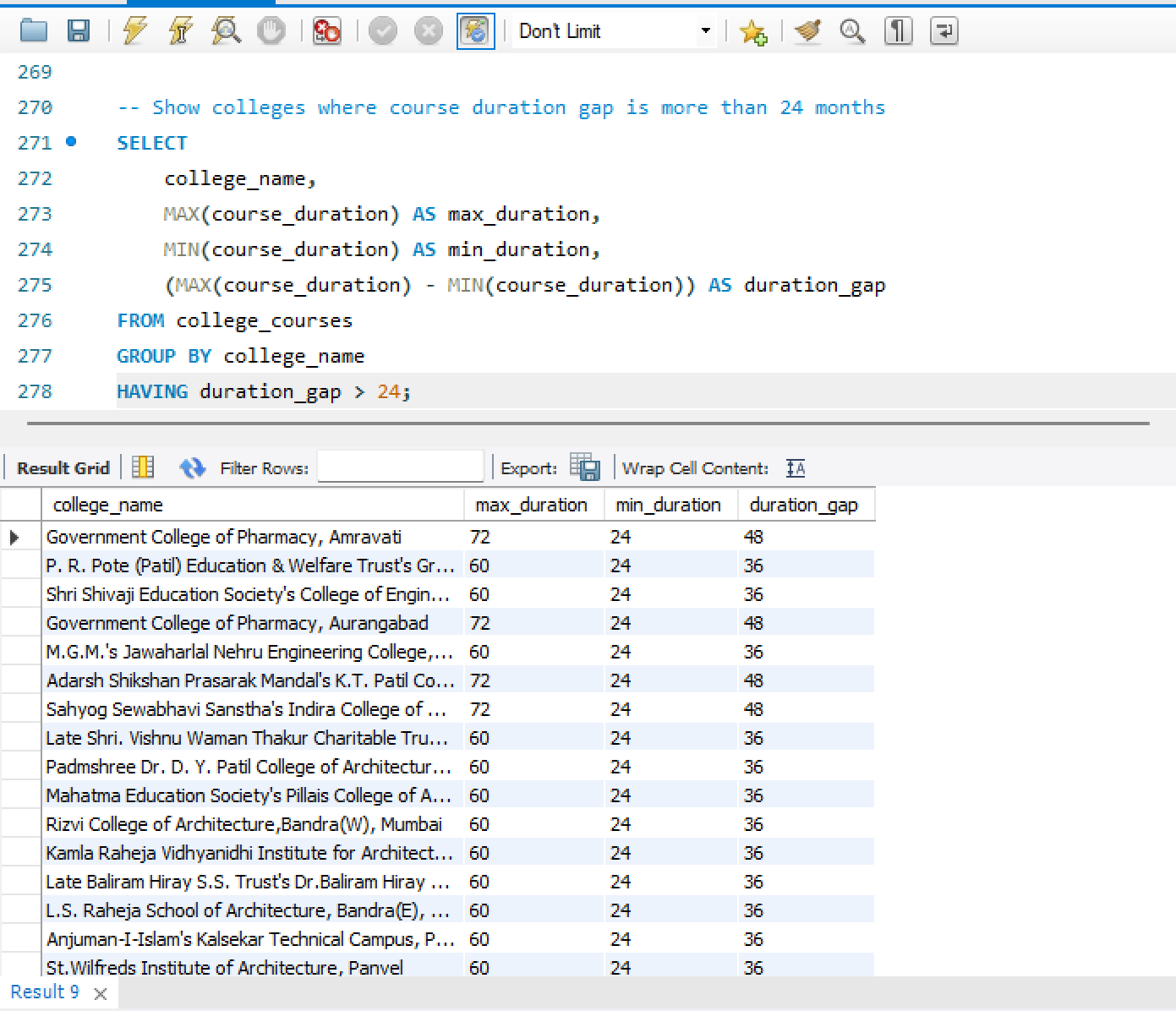
**MIN**(course\_duration) **AS** min\_duration,

(**MAX**(course\_duration) - **MIN**(course\_duration)) **AS** duration\_gap

**FROM** college\_courses

**GROUP** **BY** college\_name

**HAVING** duration\_gap > 24;



Some colleges offer both very short and long-duration programs, which shows they provide a mix of certification, diploma, and full degree courses.

Q9. Show the cumulative number of professional courses offered by each university sorted alphabetically.

**Query –**

-- List cumulative count of professional courses by university (A to Z)

**SELECT**

university,

**COUNT**(\*) **AS** professional\_course\_count,

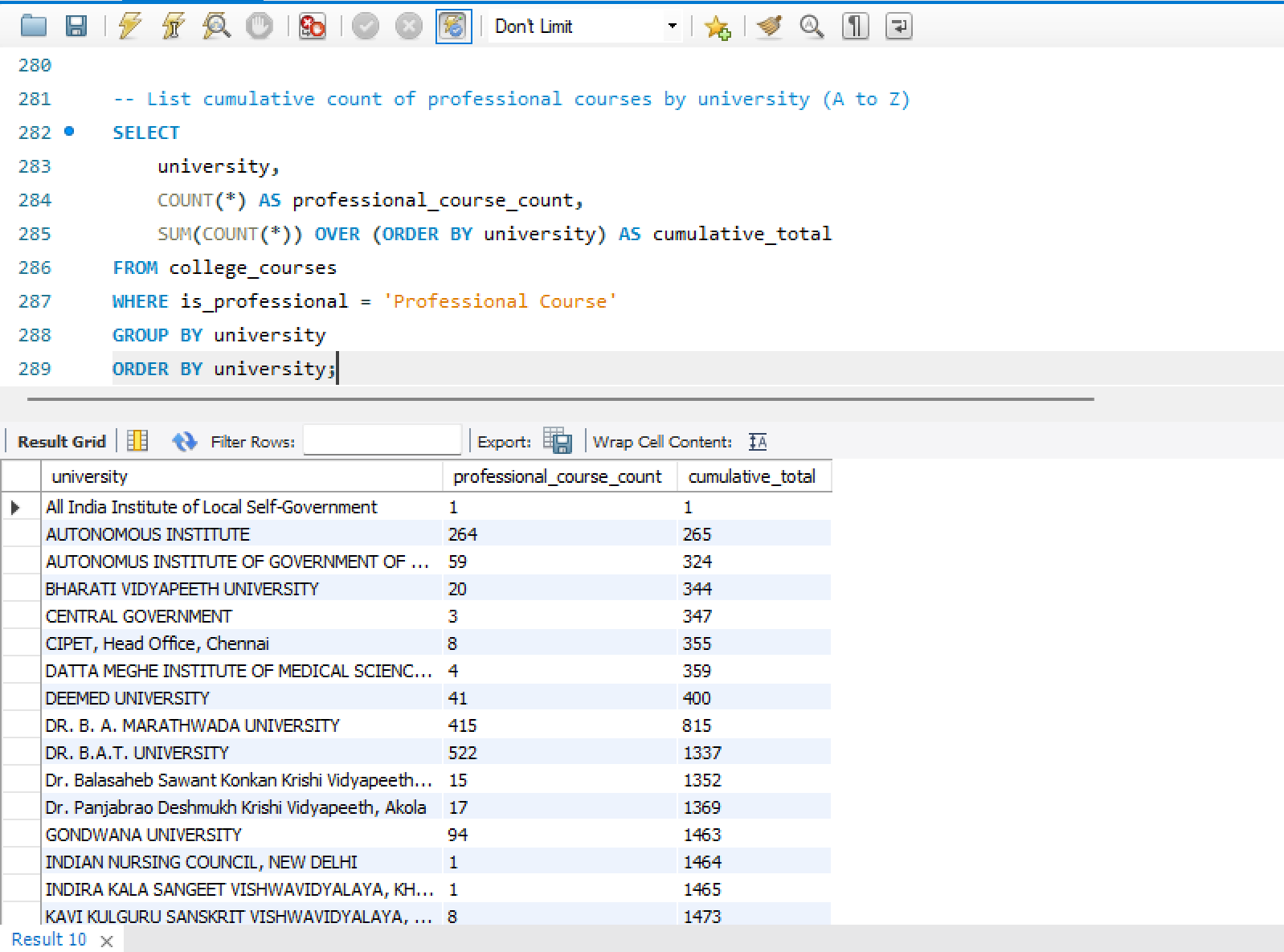
**SUM**(**COUNT**(\*)) **OVER** (**ORDER** **BY** university) **AS** cumulative\_total

**FROM** college\_courses

**WHERE** is\_professional = 'Professional Course'

**GROUP** **BY** university

**ORDER** **BY** university;



From the cumulative result, we can see the progressive count of professional courses across universities, with some contributing significantly more than others.

Q10. Using a self-join or CTE, find colleges offering more than one course category.

**Query –**

-- Use CTE to count distinct course categories per college

**WITH** category\_count **AS** (

**SELECT**

college\_name,

**COUNT**(**DISTINCT** course\_category) **AS** total\_categories

**FROM** college\_courses

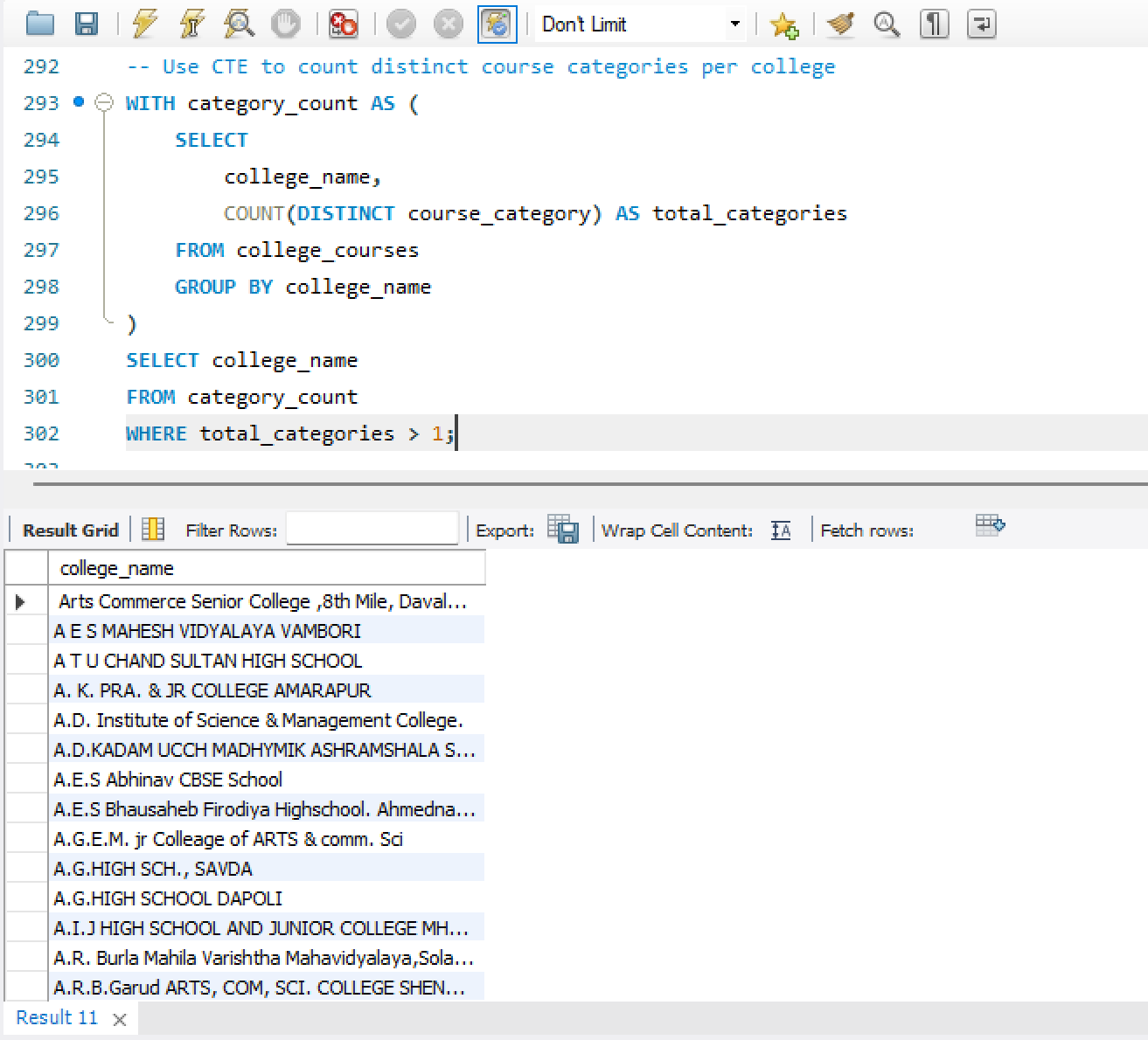
**GROUP** **BY** college\_name

)

**SELECT** college\_name

**FROM** category\_count

**WHERE** total\_categories > 1;



Many colleges offer more than one course category, meaning they don’t limit themselves to just one academic stream and cater to broader interests.

Q11. Create a temporary table (CTE) that includes average duration of courses by district and use it to list talukas where the average course duration is above the district average.

**Query –**

-- Step 1: Calculate average course duration for each district

**WITH** district\_avg\_duration **AS** (

**SELECT**

district,

**AVG**(course\_duration) **AS** district\_avg

**FROM** college\_courses

**GROUP** **BY** district

),

-- Step 2: Calculate average course duration for each taluka

taluka\_avg\_duration **AS** (

**SELECT**

district,

taluka,

**AVG**(course\_duration) **AS** taluka\_avg

**FROM** college\_courses

**GROUP** **BY** district, taluka

)

-- Step 3: Compare taluka average with district average

**SELECT**

t.taluka,

t.district,

**ROUND**(t.taluka\_avg, 2) **AS** taluka\_avg\_duration,

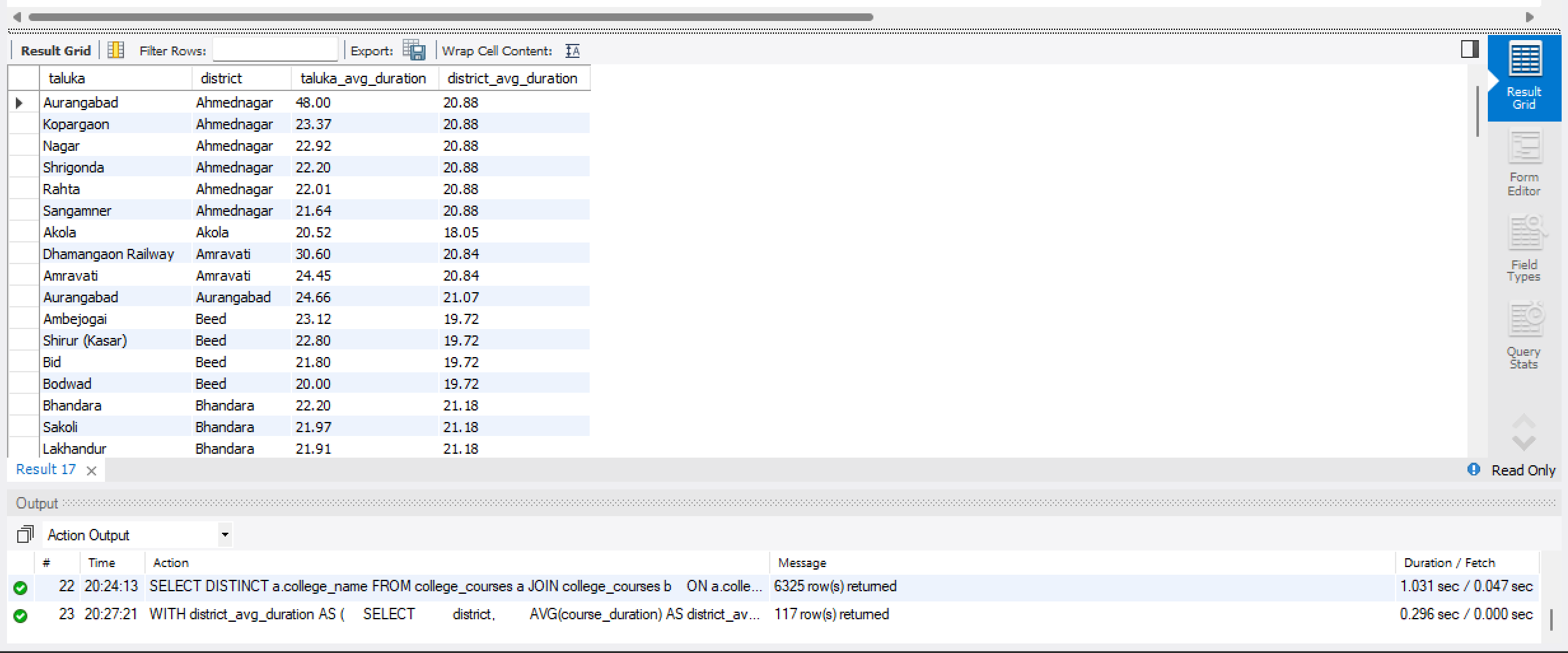
**ROUND**(d.district\_avg, 2) **AS** district\_avg\_duration

**FROM** taluka\_avg\_duration t

**JOIN** district\_avg\_duration d **ON** t.district = d.district

**WHERE** t.taluka\_avg > d.district\_avg

**ORDER** **BY** t.district, t.taluka\_avg **DESC**;



Talukas like Aurangabad and Kopargaon show higher average course durations than their district averages, indicating more advanced or longer programs.

Q12. Create a new column classifying course duration as:

* Short (< 12 months)
* Medium (12-36 months)
* Long (> 36 months)

Then count the number of each duration type per course category.

**Query –**

-- Classify courses by duration and count them per course category

**SELECT**

course\_category,

**CASE**

**WHEN** course\_duration < 12 **THEN** 'Short'

**WHEN** course\_duration **BETWEEN** 12 **AND** 36 **THEN** 'Medium'

**WHEN** course\_duration > 36 **THEN** 'Long'

**ELSE** 'Unknown'

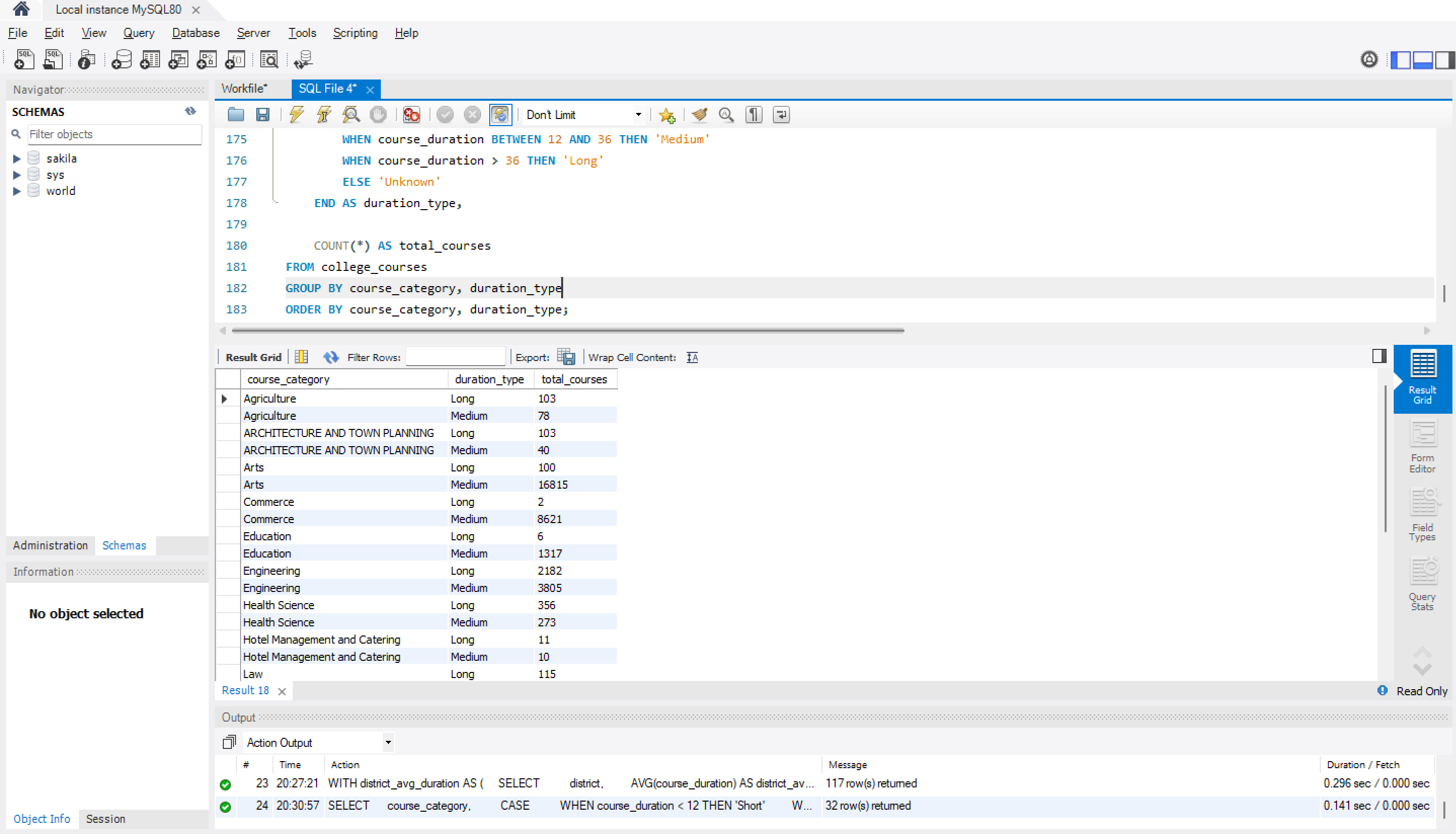
**END** **AS** duration\_type,

**COUNT**(\*) **AS** total\_courses

**FROM** college\_courses

**GROUP** **BY** course\_category, duration\_type

**ORDER** **BY** course\_category, duration\_type;



Most course categories have a majority of medium-duration programs (12–36 months), while long courses are fewer and mostly in specialized categories.

Q13. Extract only the course specialization from Course Name. (e.g., from "Bachelor of Engineering (B. E.) - Electrical", extract "Electrical")

**Query –**

-- Extract specialization from course\_name after the last dash (-)

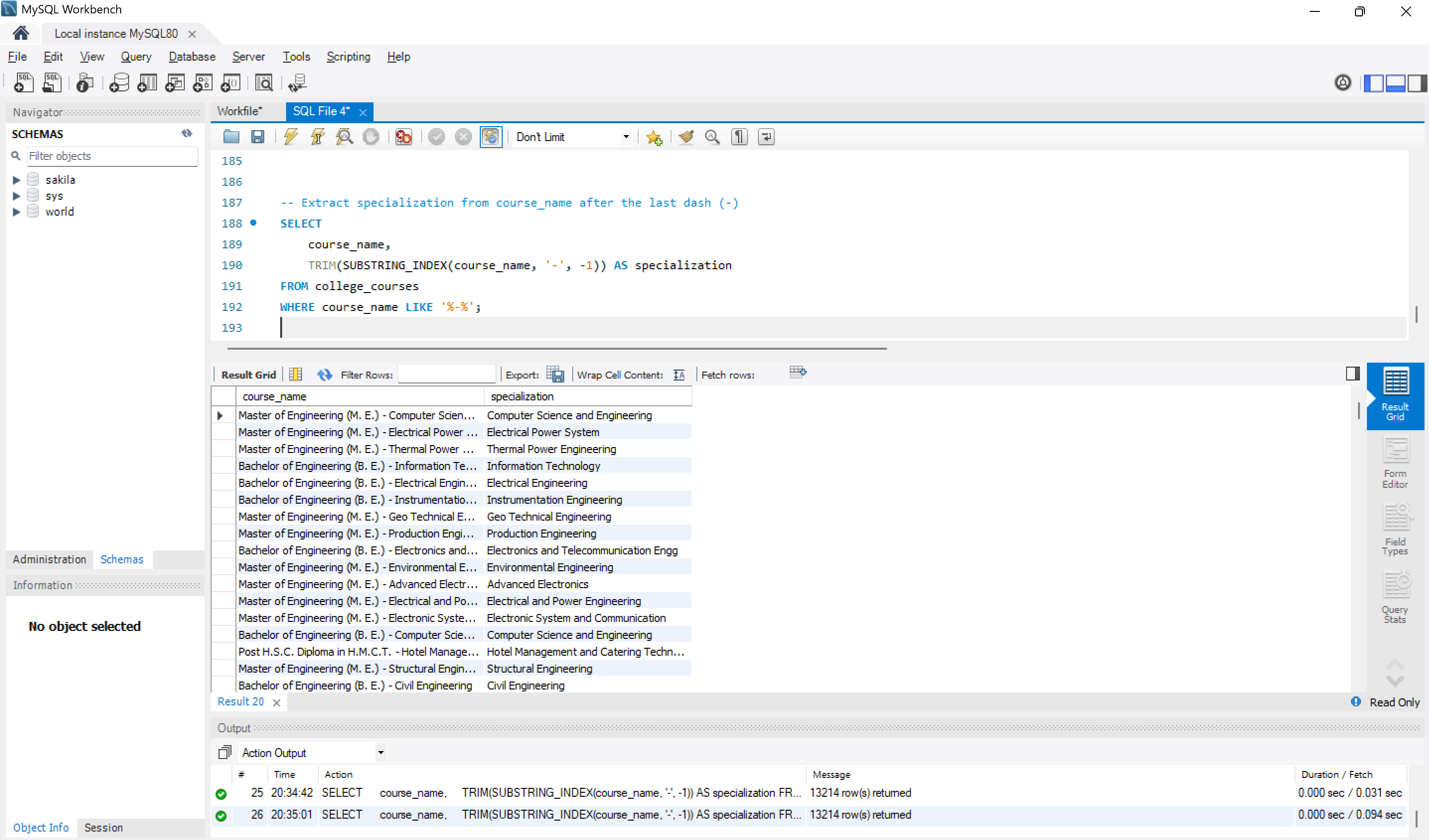
**SELECT**

course\_name,

**TRIM**(**SUBSTRING**\_**INDEX**(course\_name, '-', -1)) **AS** specialization

**FROM** college\_courses

**WHERE** course\_name **LIKE** '%-%';



Specializations like Electrical, Computer Science, and Civil are extracted from the course names, which helps in grouping courses more meaningfully.

Q14. Count how many courses include the word Engineering in the name.

**Query –**

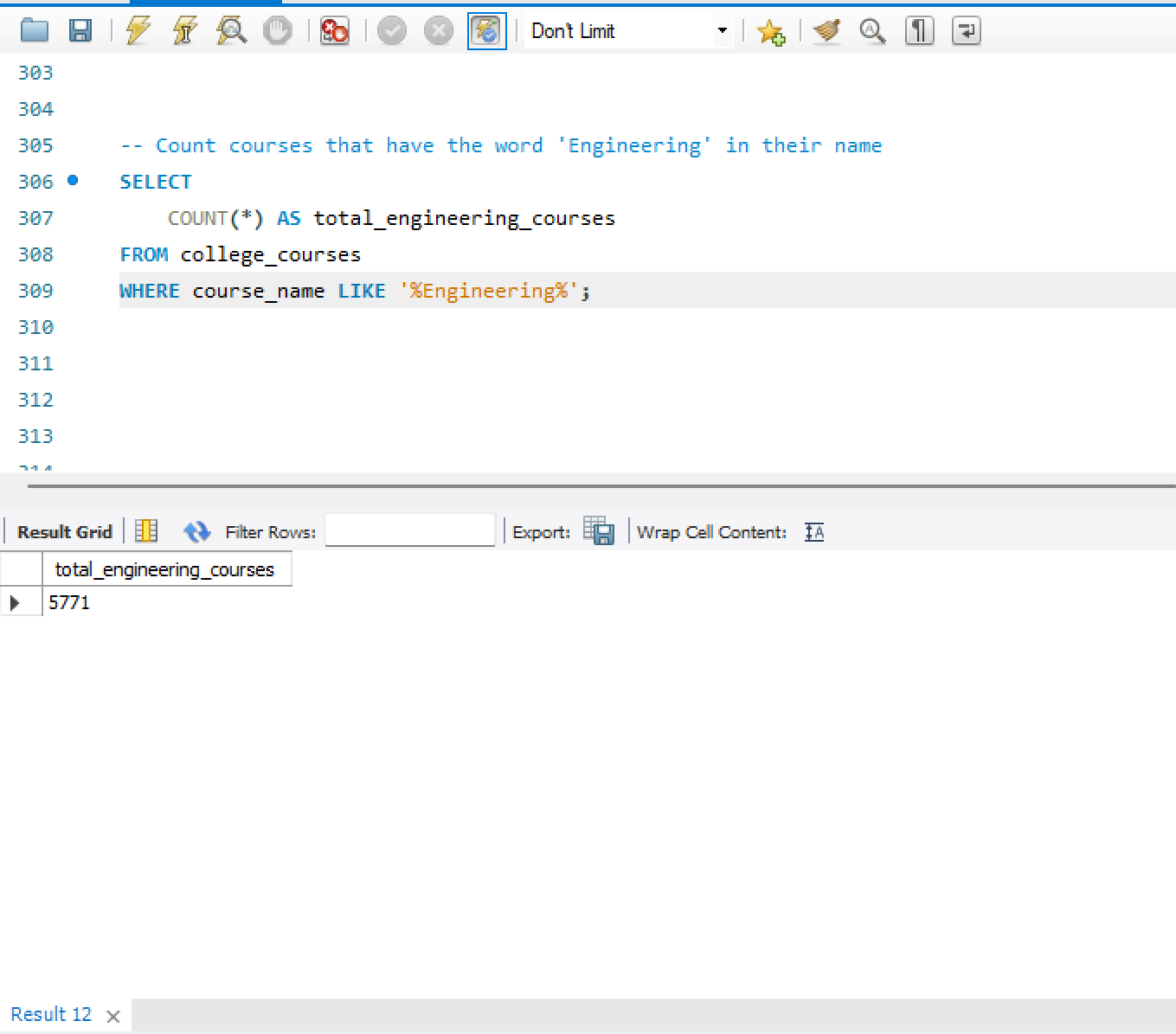
-- Count courses that have the word 'Engineering' in their name

**SELECT**

**COUNT**(\*) **AS** total\_engineering\_courses

**FROM** college\_courses

**WHERE** course\_name **LIKE** '%Engineering%';



A total of 5771 courses have the word "Engineering" in their name, confirming the high number of technical programs in the dataset.

Q15. List all unique combinations of Course Name, Course Type, and Course Category.

**Query –**

-- Show distinct combinations of course name, type, and category

**SELECT** **DISTINCT**

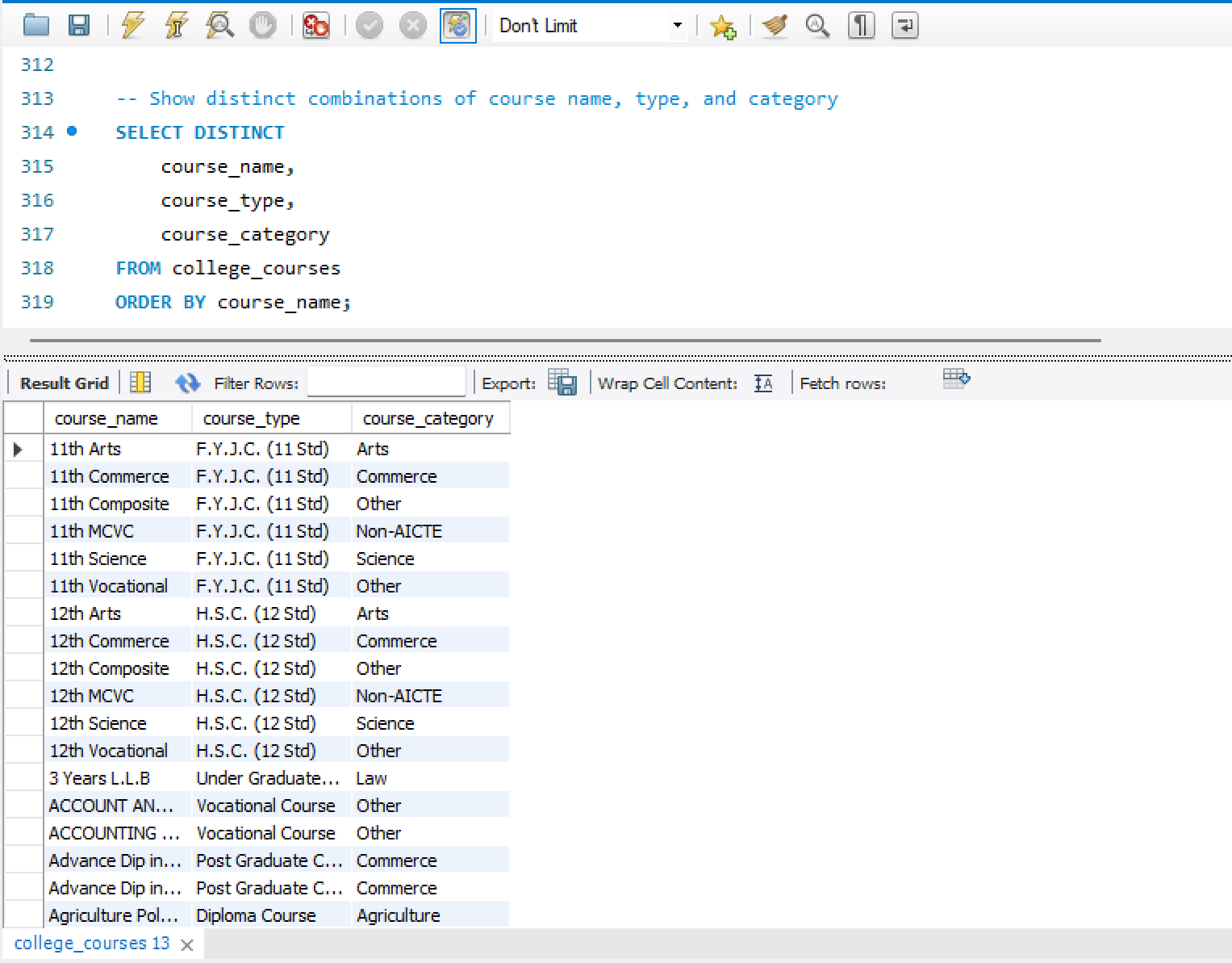
course\_name,

course\_type,

course\_category

**FROM** college\_courses

**ORDER** **BY** course\_name;



The output lists various unique combinations of course name, type, and category — useful to analyze the academic structure and duplication.

Q16. Write a query to get all courses that are not offered by any Government college.

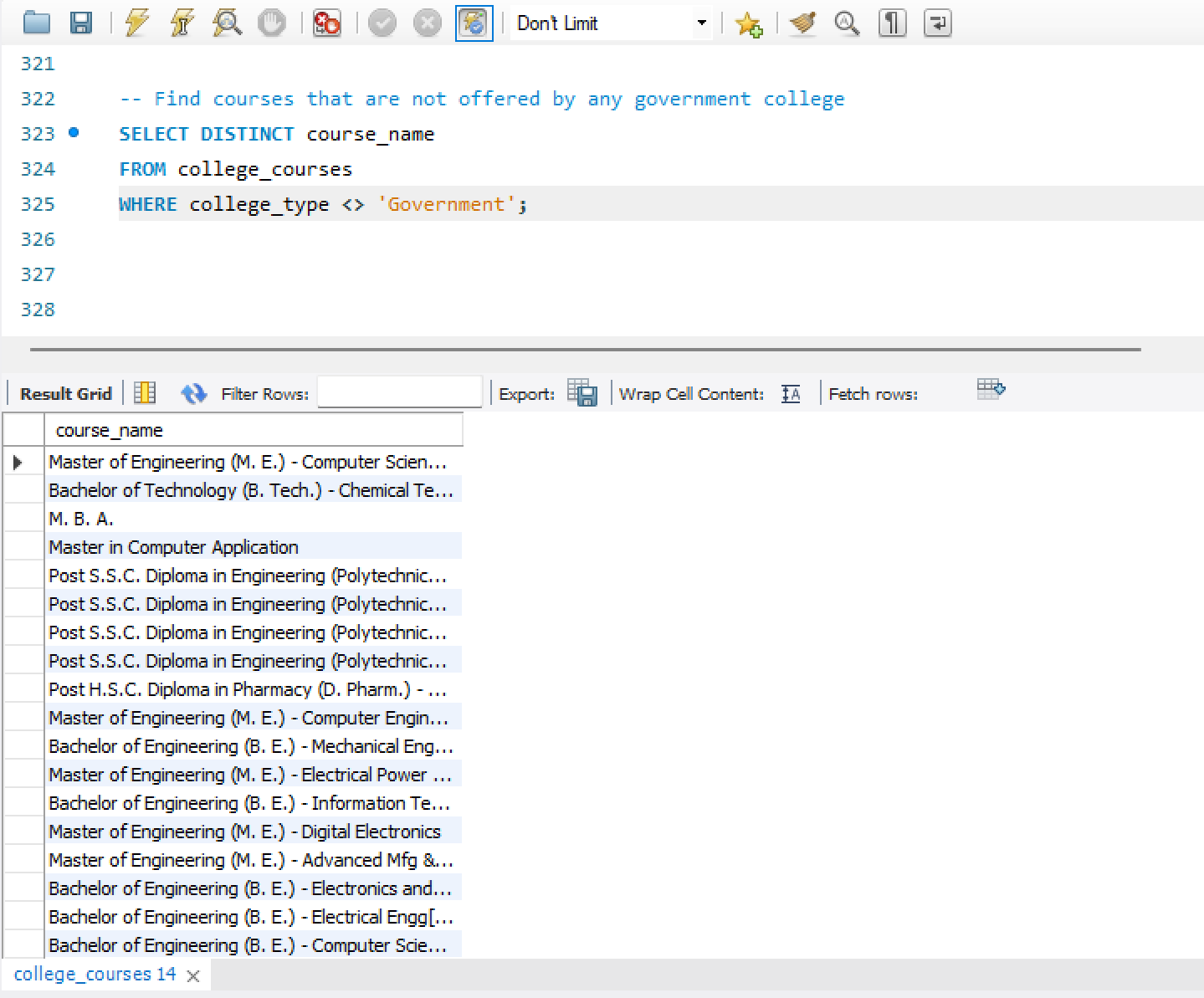
**Query –**

-- Find courses that are not offered by any government college

**SELECT** **DISTINCT** course\_name

**FROM** college\_courses

**WHERE** college\_type <> 'Government';



The query successfully filtered out courses that are not offered in any government college — mostly niche, short-term, or private-only courses.

Q17. Find the university that has the second-highest number of aided courses.

**Query –**

-- Find the university with the second-highest number of aided courses

**SELECT** university, total\_aided\_courses

**FROM** (

**SELECT**

university,

**COUNT**(\*) **AS** total\_aided\_courses,

**RANK**() **OVER** (**ORDER** **BY** **COUNT**(\*) **DESC**) **AS** rank\_num

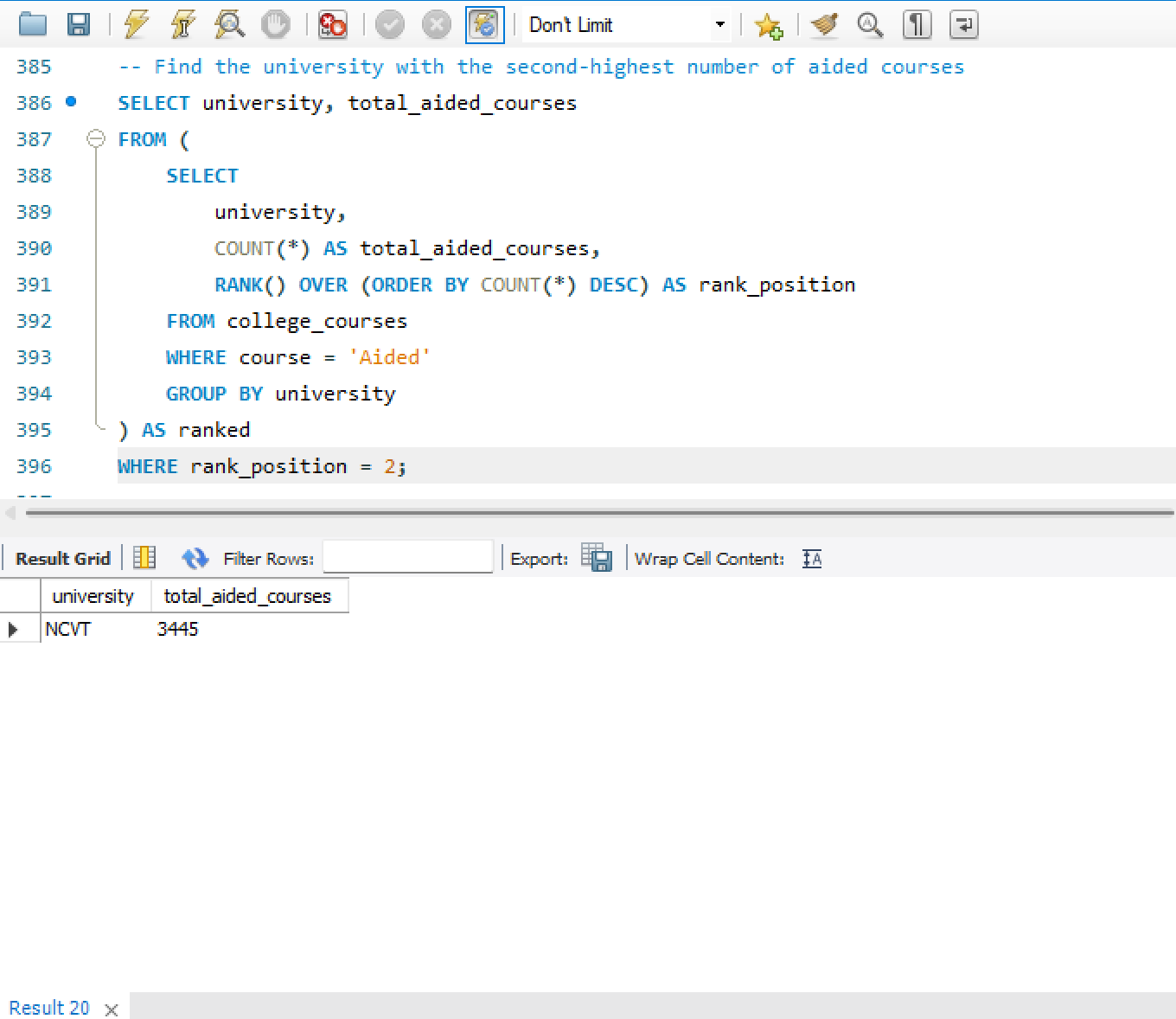
**FROM** college\_courses

**WHERE** course = 'Aided'

**GROUP** **BY** university

) **AS** ranked\_universities

**WHERE** rank\_num = 2;



NCVT university with the second-highest number of aided courses was identified clearly, which helps compare institutional support types.

Q18. Show courses whose durations are above the median course duration.

**Query –**

-- Find courses that have duration above the median

**SELECT** course\_name, course\_duration

**FROM** college\_courses

**WHERE** course\_duration > (

**SELECT** course\_duration

**FROM** (

**SELECT** course\_duration,

**ROW**\_**NUMBER**() **OVER** (**ORDER** **BY** course\_duration) **AS** row\_num,

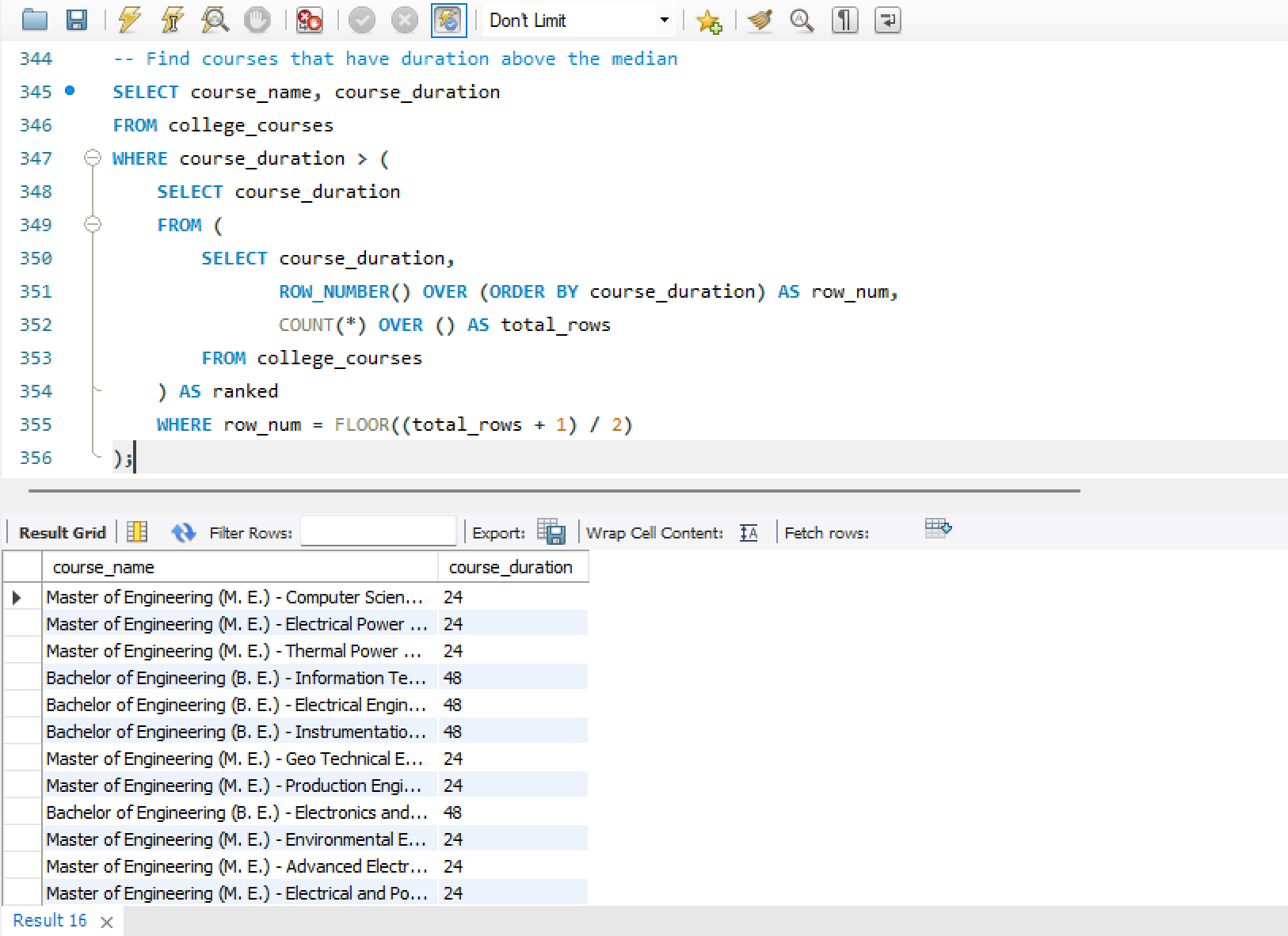
**COUNT**(\*) **OVER** () **AS** total\_rows

**FROM** college\_courses

) **AS** ranked

**WHERE** row\_num = **FLOOR**((total\_rows + 1) / 2)

);



Courses with durations above the median are mostly higher-level programs, such as degrees and PG courses.

Q19. For each University, find the percentage of unaided courses that are professional.

**Query –**

-- Calculate percentage of unaided courses that are professional for each university

**SELECT**

university,

**ROUND**(

**SUM**(**CASE** **WHEN** is\_professional = 'Professional Course' **THEN** 1 **ELSE** 0 **END**) \* 100.0

/ **COUNT**(\*), 2

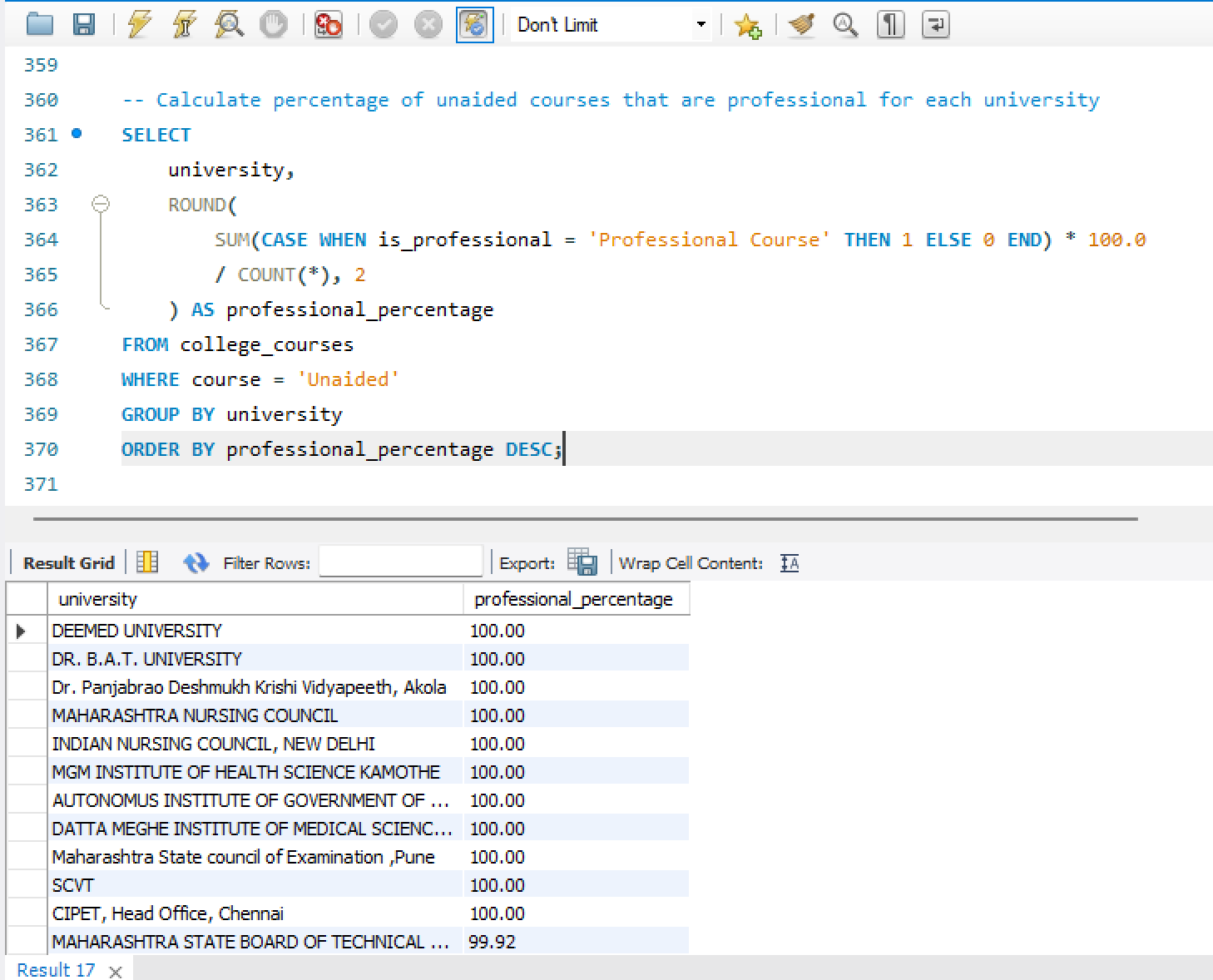
) **AS** professional\_percentage

**FROM** college\_courses

**WHERE** course = 'Unaided'

**GROUP** **BY** university

**ORDER** **BY** professional\_percentage **DESC**;



Some universities have a high percentage of professional courses among their unaided offerings, showing a focus on job-oriented programs.

Q20. Determine which Course Category has the highest average course duration and display the top 3.

**Query—**

-- Get top 3 course categories with highest average duration

**SELECT**

course\_category,

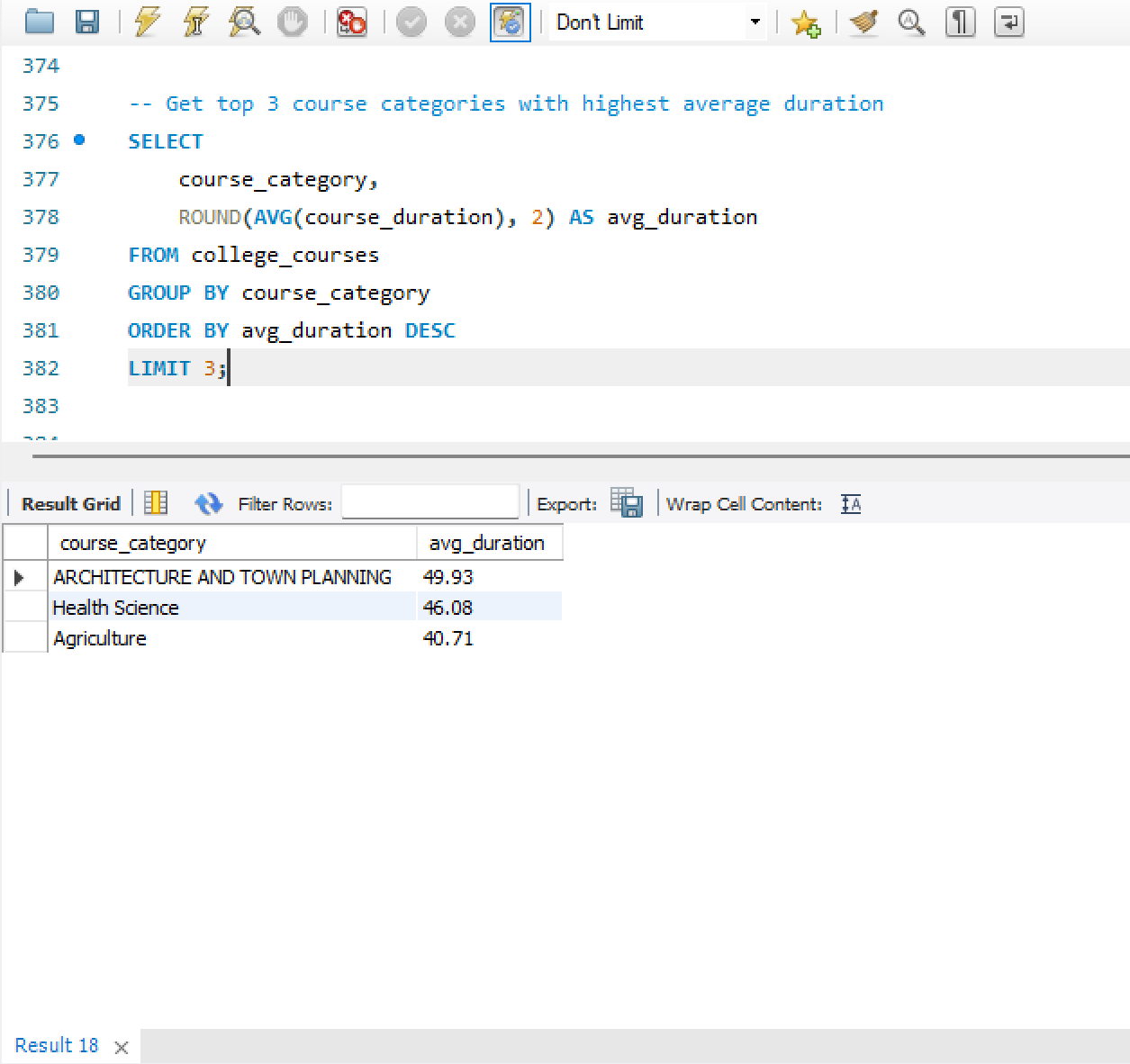
**ROUND**(**AVG**(course\_duration), 2) **AS** avg\_duration

**FROM** college\_courses

**GROUP** **BY** course\_category

**ORDER** **BY** avg\_duration **DESC**

**LIMIT** 3;



Architecture and Town Planning, Health Science , and Agriculture have the highest average durations, reflecting their depth and complexity.

# **CONCLUSION**

"This project helped analyze the distribution of professional and non-professional courses across various universities and colleges in Maharashtra. It provided insights into course duration patterns, popular categories, and the availability of aided/unaided programs."