PSYLIQ

DATA ANALYST INTERNSHIP

DIABETES PREDICTION ASSESSMENT

INTRODUCTION

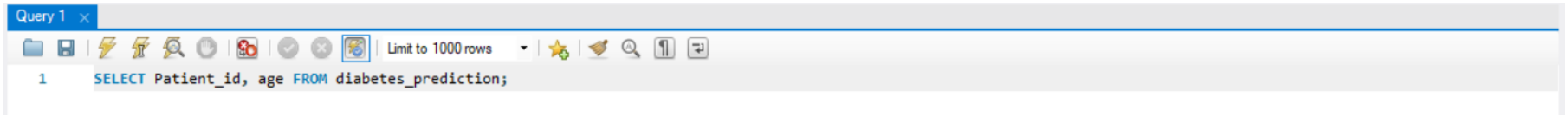
HELLO AND WELCOME! TO OUR SQL PROJECT REPORT ON DIABETES PREDICTION. IN THIS PROJECT, WE EMBARK THE EXCITING WORLD OF DATA SCIENCE AND HEALTHCARE, COMBINING THE POWER OF SQL WITH THE CRUCIAL MISSION OF DIABETES PREDICTION.

WE'LL BE WORKING WITH DATASET THAT INCLUDES GENDER, AGE, HYPERTENSION, HEART DISEASE, SMOKING HISTORY, BMI, HBA1C LEVELS, BLOOD GLUCOSE LEVELS, AND PRESENCE OR ABSENCE OF DIABETES DURING THIS PROJECT. THIS REAL WORLD DATASET REFLECTS THE COMPLEXITIES OF HEALTHCARE DATA AND OFFERS AN EXCELLENT ENVIRONMENT FOR HONING YOUR SQL SKILLS.

TASK

1. Retrieve the Patient_id and ages of all patients.

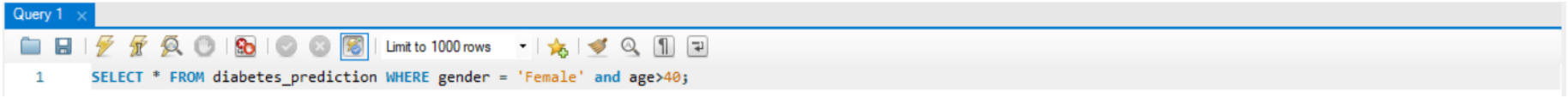
Query



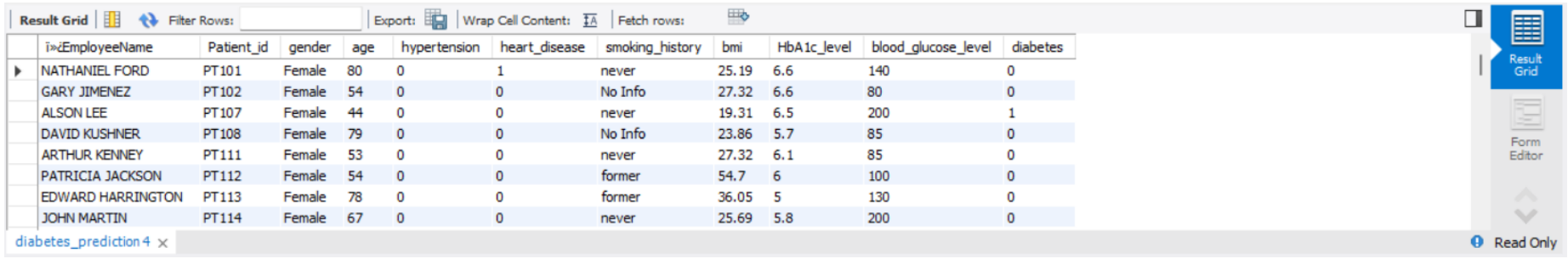
Results

2. Select all female patients who are older than 40.

Query



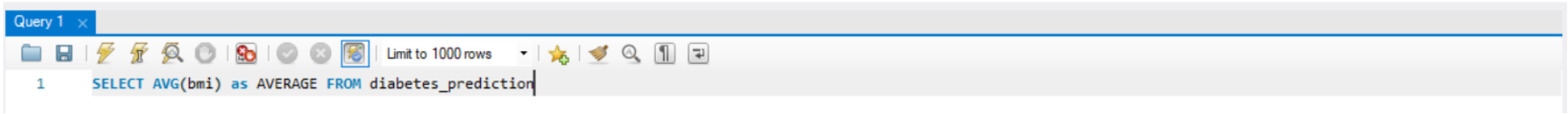
Results



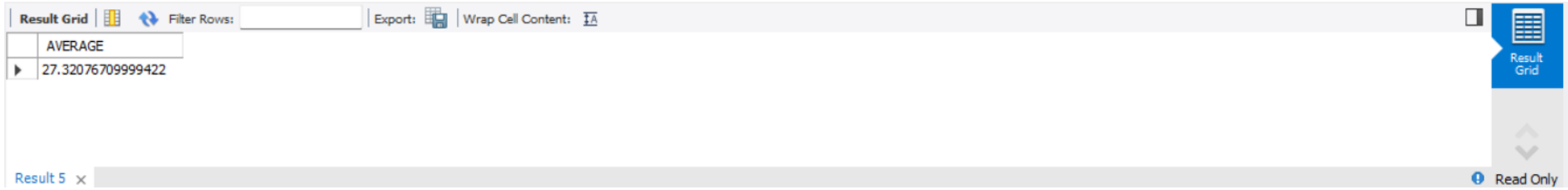
ï»¿EmployeeName	Patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
NATHANIEL FORD	PT101	Female	80	0	1	never	25.19	6.6	140	0
GARY JIMENEZ	PT102	Female	54	0	0	No Info	27.32	6.6	80	0
ALSON LEE	PT107	Female	44	0	0	never	19.31	6.5	200	1
DAVID KUSHNER	PT108	Female	79	0	0	No Info	23.86	5.7	85	0
ARTHUR KENNEY	PT111	Female	53	0	0	never	27.32	6.1	85	0
PATRICIA JACKSON	PT112	Female	54	0	0	former	54.7	6	100	0
EDWARD HARRINGTON	PT113	Female	78	0	0	former	36.05	5	130	0
JOHN MARTIN	PT114	Female	67	0	0	never	25.69	5.8	200	0

3. Calculate the average BMI of patients.

Query



Results



AVERAGE
27.32076709999422

4. List patients in descending order of blood glucose levels.

Query

```
1   SELECT * FROM diabetes_prediction ORDER BY blood_glucose_level DESC;
```

Results

ï»¿EmployeeName	Patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Michelle D McGee	PT98852	Male	79	0	0	ever	27.32	7.5	300	1
Lawrence Shum	PT98855	Male	43	0	0	former	48.56	6.8	300	1
Seth I Rubenstein	PT98911	Female	60	0	0	current	40.18	9	300	1
Philip Tran	PT99008	Male	69	0	0	never	31.56	7	300	1
Gilbert J Fragoso	PT99638	Female	67	1	0	ever	34.3	5.7	300	1

5. Find patients who have hypertension and diabetes.

Query

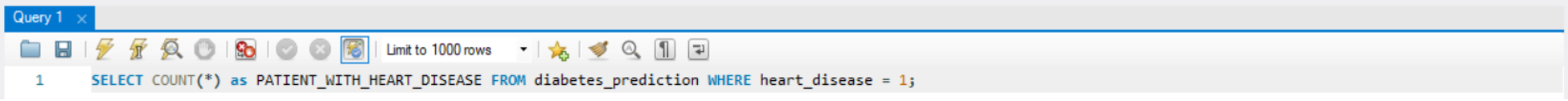
```
1   SELECT * FROM diabetes_prediction WHERE hypertension = 1 AND diabetes = 1;
```

Results

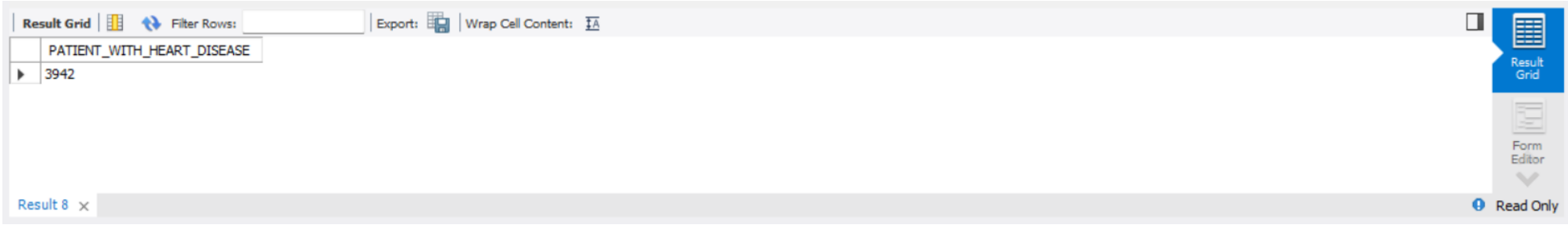
ï»¿EmployeeName	Patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
JONES WONG	PT139	Male	50	1	0	current	27.32	5.7	260	1
PATRIC STEELE	PT205	Female	80	1	0	never	27.32	6.8	280	1
ARTHUR STELLINI	PT343	Male	57	1	1	not current	27.77	6.6	160	1
CHAD LAW	PT355	Male	63	1	0	ever	35.06	5.8	200	1
CATHERINE JAMES	PT451	Female	52	1	0	never	50.3	6.6	155	1
JOHN HART	PT565	Male	48	1	0	current	36.12	6.8	140	1

6. Determine the number of patients with heart disease.

Query

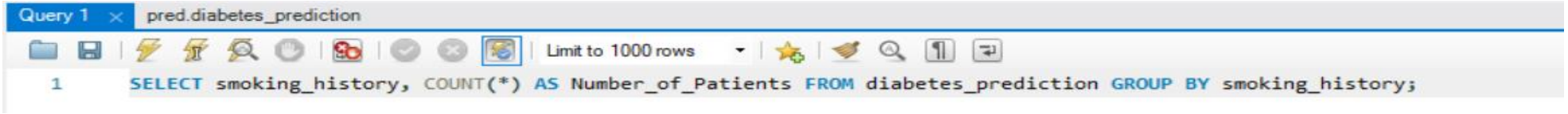


Results

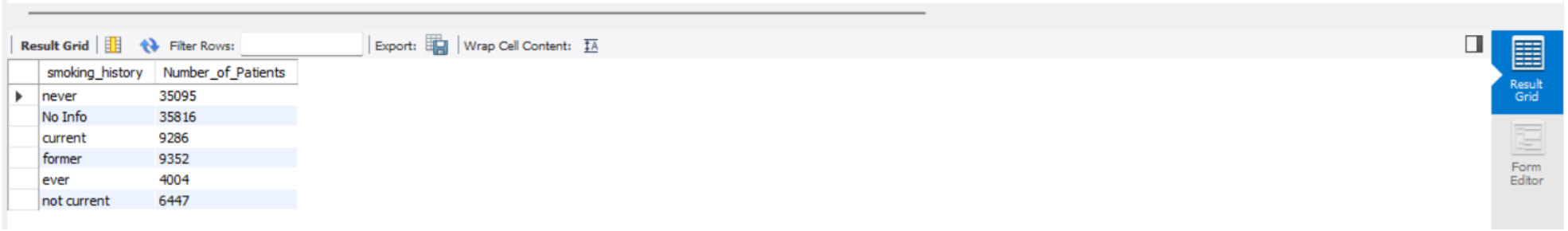


7. Group patients by smoking history and count how many smokers and non-smokers there are.

Query

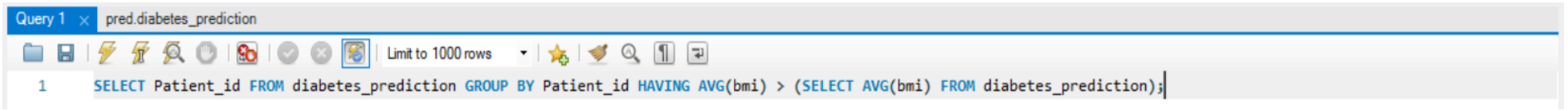


Results



8. Retrieve the Patient IDs of patients who have a BMI greater than the average BMI.

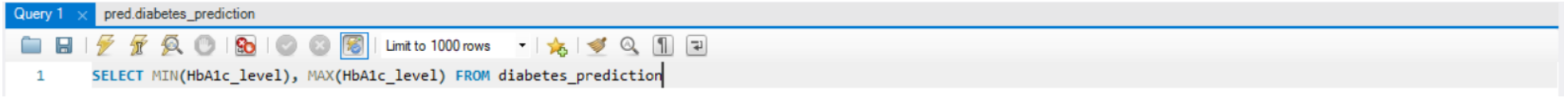
Query



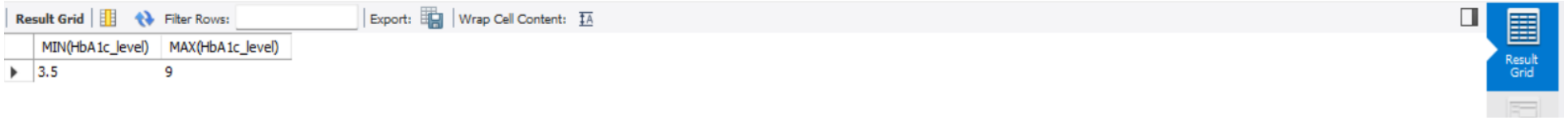
Results

9. Find the patient with the highest HbA1c level and the patient with the lowest HbA1clevel.

Query



Results



10. Calculate the age of patients in years (assuming the current date as of now).

Query

```
SELECT patient_id, ABS((age - YEAR(NOW()))) AS year_of_birth FROM diabetes_prediction
```

Results

patient_id	year_of_birth
PT101	1943
PT102	1969
PT103	1995
PT104	1987
PT105	1947
PT106	2003
PT107	1979
PT108	1944
PT109	1981

11. Rank patients by blood glucose level within each gender group.

Query

```
SELECT patient_id, gender, ROW_NUMBER() OVER(ORDER BY blood_glucose_level DESC, gender DESC) AS Patients_Rank FROM diabetes_prediction
```

Results

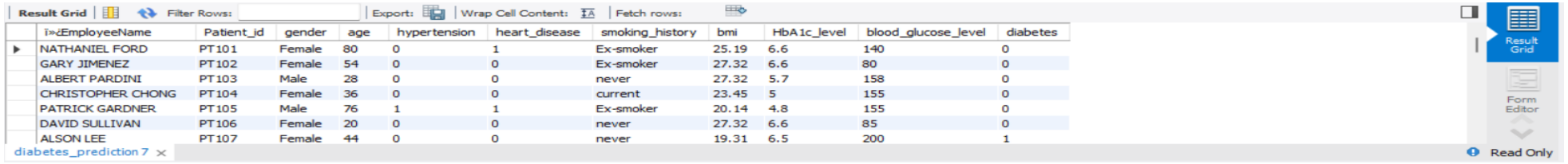
patient_id	gender	Patients_Rank
PT96269	Male	1
PT97419	Male	2
PT98461	Male	3
PT99663	Male	4
PT98855	Male	5
PT98852	Male	6
PT99968	Male	7
PT99008	Male	8
PT99809	Male	9

12. Update the smoking history of patients who are older than 50 to "Ex-smoker."

Query

```
SET SQL_SAFE_UPDATES = 0;
UPDATE diabetes_prediction SET smoking_history = 'Ex-smoker' WHERE age > 50;
    SELECT * FROM diabetes_prediction;
```

Results



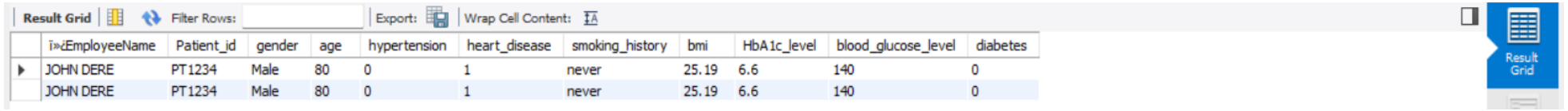
EmployeeName	Patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
NATHANIEL FORD	PT101	Female	80	0	1	Ex-smoker	25.19	6.6	140	0
GARY JIMENEZ	PT102	Female	54	0	0	Ex-smoker	27.32	6.6	80	0
ALBERT PARDINI	PT103	Male	28	0	0	never	27.32	5.7	158	0
CHRISTOPHER CHONG	PT104	Female	36	0	0	current	23.45	5	155	0
PATRICK GARDNER	PT105	Male	76	1	1	Ex-smoker	20.14	4.8	155	0
DAVID SULLIVAN	PT106	Female	20	0	0	never	27.32	6.6	85	0
ALSON LEE	PT107	Female	44	0	0	never	19.31	6.5	200	1

13. Insert a new patient into the database with sample data.

Query

```
INSERT INTO diabetes_prediction
(EmployeeName, Patient_id, gender, age, hypertension, heart_disease, smoking_history, bmi, HbA1c_level, blood_glucose_level, diabetes)
VALUES ('JOHN DERE', 'PT1234', 'Male', 80, 0, 1, 'never', 25.19, 6.6, 140, 0);
SELECT * FROM diabetes_prediction WHERE EmployeeName = 'JOHN DERE';
```

Results



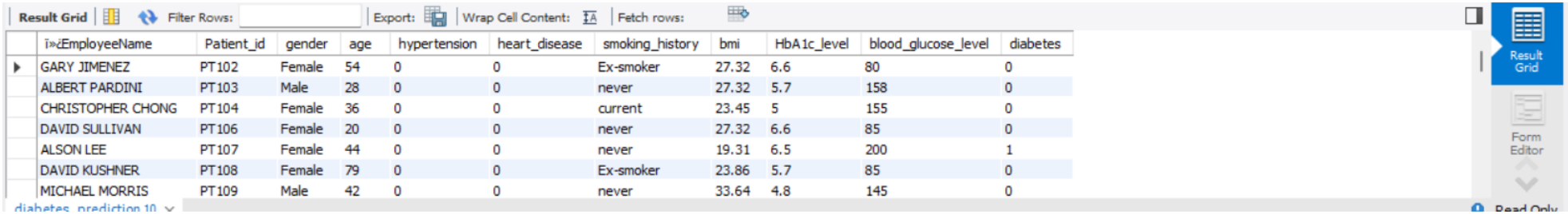
EmployeeName	Patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
JOHN DERE	PT1234	Male	80	0	1	never	25.19	6.6	140	0
JOHN DERE	PT1234	Male	80	0	1	never	25.19	6.6	140	0

14. Delete all patients with heart disease from the database.

Query

```
SET SQL_SAFE_UPDATES = 0;
DELETE FROM diabetes_prediction WHERE heart_disease = 1;
SELECT * FROM diabetes_prediction;
```

Results



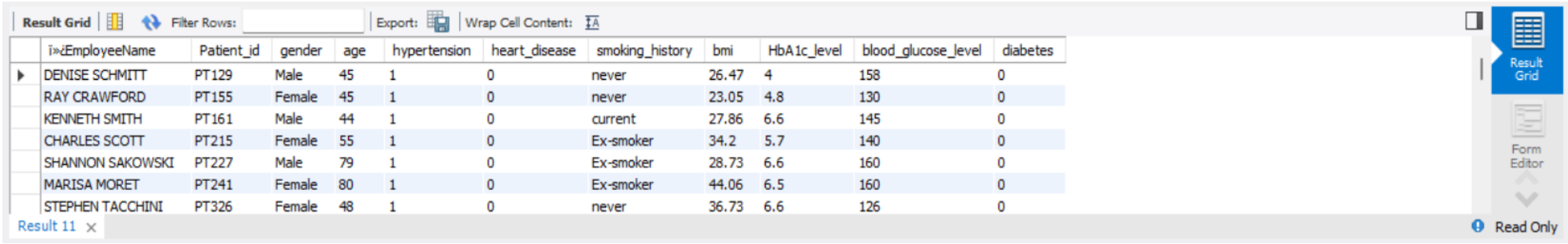
ï»¿EmployeeName	Patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
GARY JIMENEZ	PT102	Female	54	0	0	Ex-smoker	27.32	6.6	80	0
ALBERT PARDINI	PT103	Male	28	0	0	never	27.32	5.7	158	0
CHRISTOPHER CHONG	PT104	Female	36	0	0	current	23.45	5	155	0
DAVID SULLIVAN	PT106	Female	20	0	0	never	27.32	6.6	85	0
ALSON LEE	PT107	Female	44	0	0	never	19.31	6.5	200	1
DAVID KUSHNER	PT108	Female	79	0	0	Ex-smoker	23.86	5.7	85	0
MICHAEL MORRIS	PT109	Male	42	0	0	never	33.64	4.8	145	0

15. Find patients who have hypertension but not diabetes using the EXCEPT operator.

Query

```
SELECT * FROM diabetes_prediction WHERE hypertension = 1 EXCEPT SELECT * FROM diabetes_prediction WHERE diabetes = 1;
```

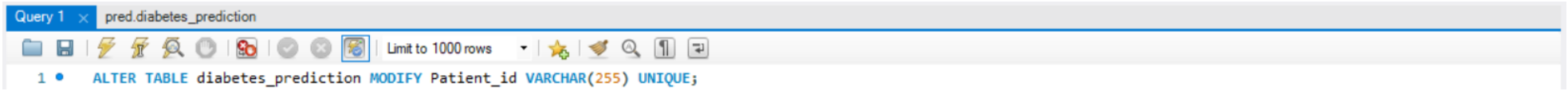
Results



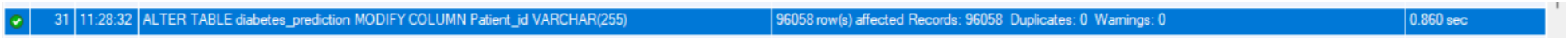
ï»¿EmployeeName	Patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
DENISE SCHMITT	PT129	Male	45	1	0	never	26.47	4	158	0
RAY CRAWFORD	PT155	Female	45	1	0	never	23.05	4.8	130	0
KENNETH SMITH	PT161	Male	44	1	0	current	27.86	6.6	145	0
CHARLES SCOTT	PT215	Female	55	1	0	Ex-smoker	34.2	5.7	140	0
SHANNON SAKOWSKI	PT227	Male	79	1	0	Ex-smoker	28.73	6.6	160	0
MARISA MORET	PT241	Female	80	1	0	Ex-smoker	44.06	6.5	160	0
STEPHEN TACCHINI	PT326	Female	48	1	0	never	36.73	6.6	126	0

16. Define a unique constraint on the "patient_id" column to ensure its values are unique.

Query

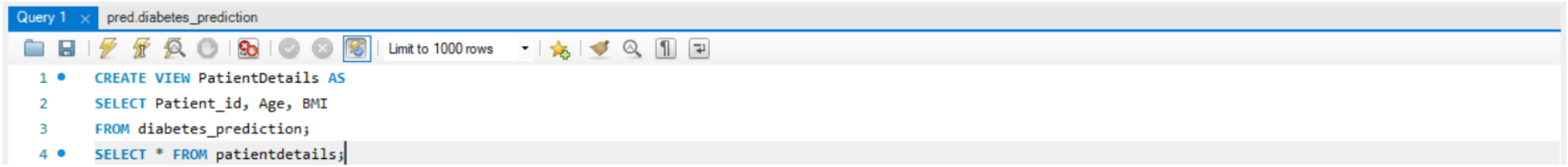


Results

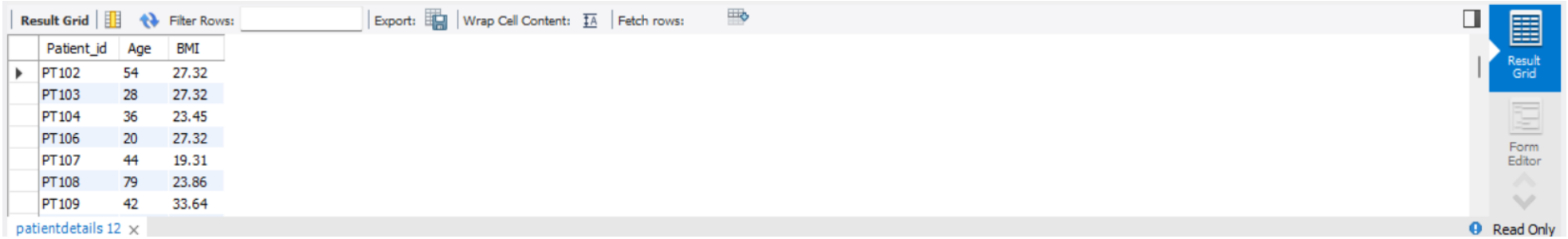


17. Create a view that displays the Patient_ids, ages, and BMI of patients.

Query



Results

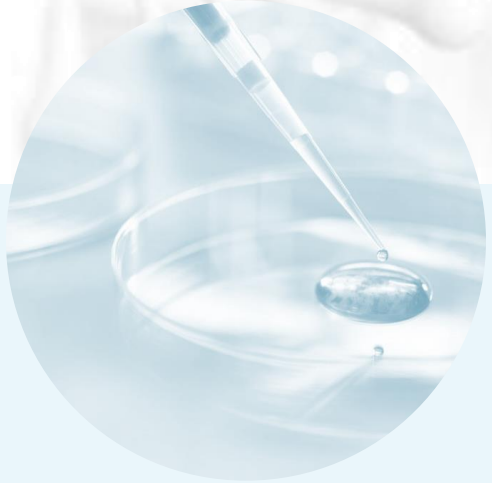


18. Suggest improvements in the database schema to reduce data redundancy and improve data integrity.

- Create indexes on frequently queried columns.
- Set default values where appropriate.
- Use constraints (e.g., NOT NULL) for data integrity.
- Use foreign keys or separate tables for shared information.
- Maintain documentation for relationships and constraints.

19. Explain how you can optimize the performance of SQL queries on this dataset.

- Replace some subqueries with JOINs or EXISTS.
- Avoid using SELECT *; only select necessary columns.
- Create indexes on frequently used columns.
- Use covering indexes.
- Encapsulate complex queries in stored procedures. This reduces network traffic and allows for better query plan caching.



THANK YOU

REHMAN KHAN
ARK.KHAN0123@GMAIL.COM | HTTPS://WWW.LINKEDIN.COM/IN/KHANREHMAN/