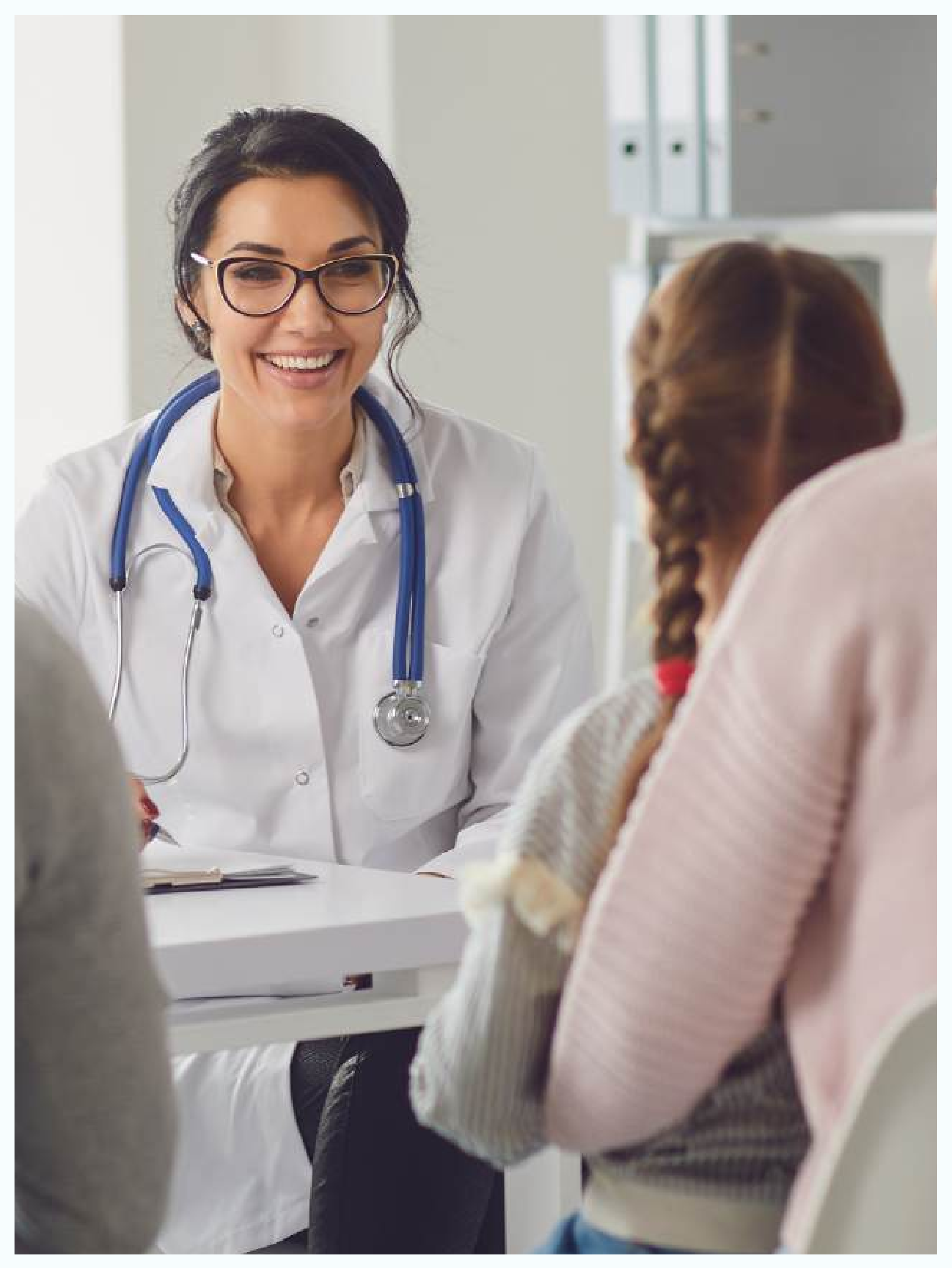
Diabetes Prediction Data Analysis using SQL

By Shravani Halaye

Objective :

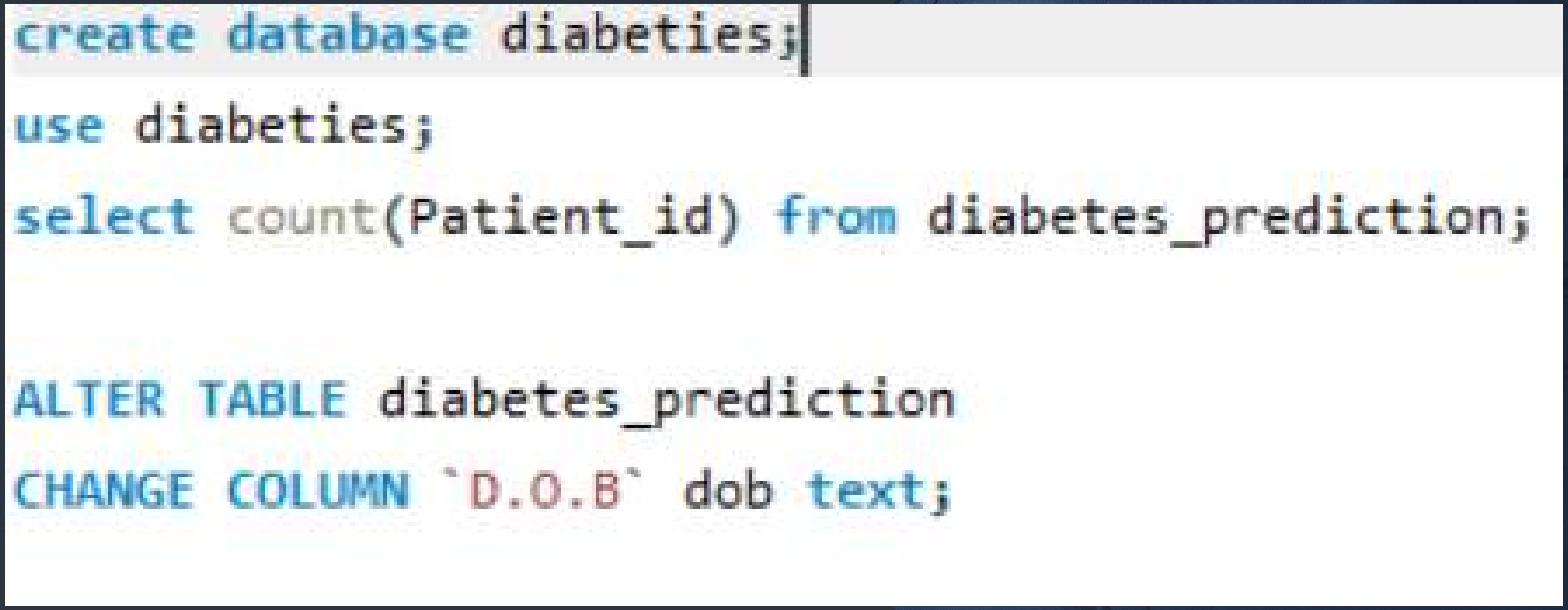
The research focuses on analysing to accurately predict the likelihood of patients developing diabetes by leveraging a structured SQL database to store and analyze health metrics and lifestyle factors.



SCHEMA

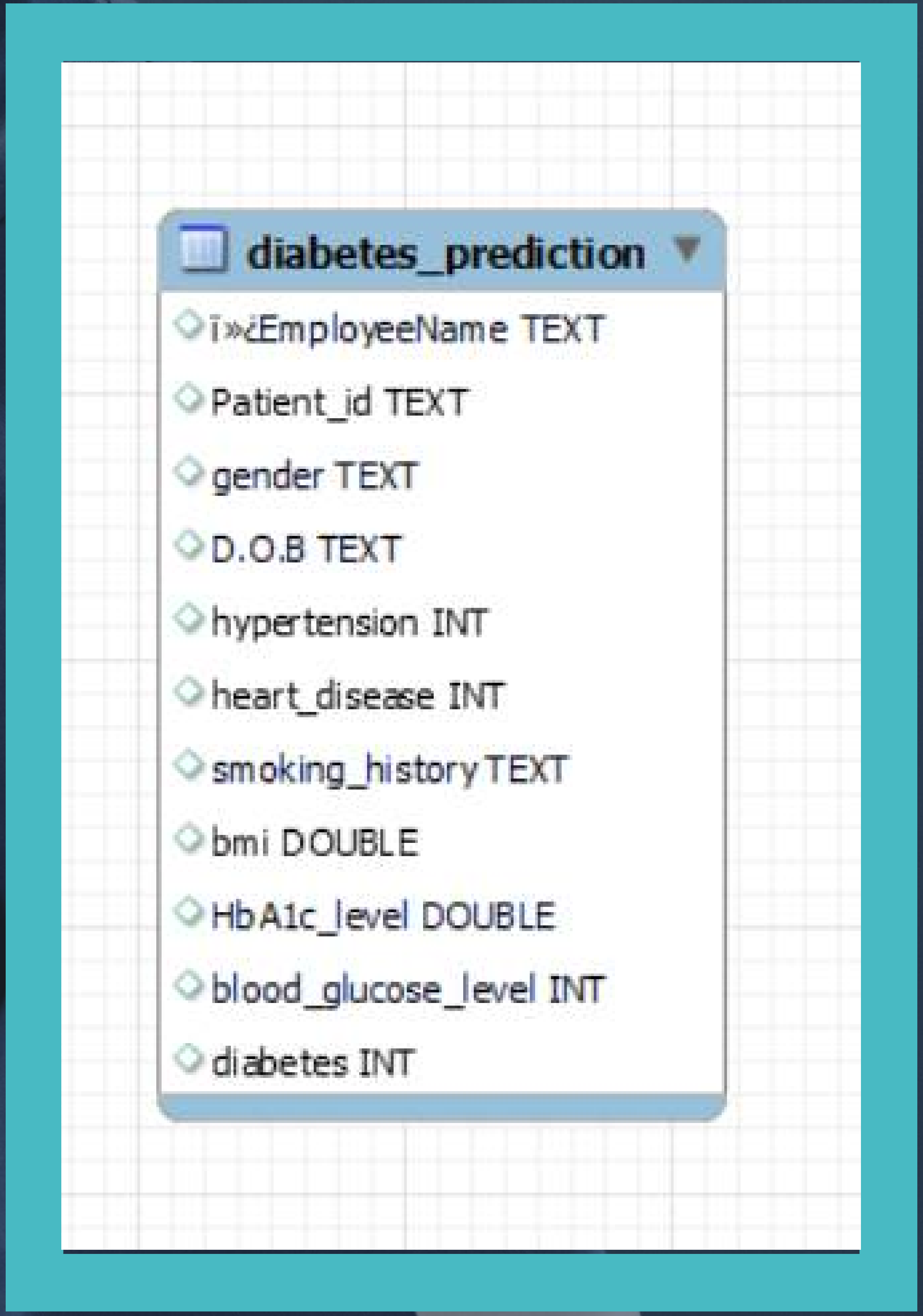
PATIENT_ID PRIMARY KEY

....



```
create database diabeties;
use diabeties;
select count(Patient_id) from diabetes_prediction;

ALTER TABLE diabetes_prediction
CHANGE COLUMN `D.O.B` dob text;
```



diabetes_prediction

- ï»¿EmployeeName TEXT
- Patient_id TEXT
- gender TEXT
- D.O.B TEXT
- hypertension INT
- heart_disease INT
- smoking_history TEXT
- bmi DOUBLE
- HbA1c_level DOUBLE
- blood_glucose_level INT
- diabetes INT

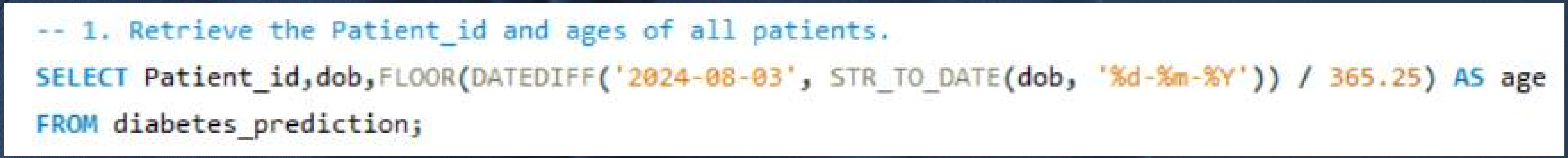
All Questions

1. Retrieve the Patient_id and ages of all patients.
2. Select all female patients who are olderthan 30.
3. Calculate the average BMI of patients.
4. List patients in descending order of blood glucose levels.
5. Find patients who have hypertension and diabetes.
6. Determine the number of patients with heart disease.
7. Group patients by smoking history and count how many smokers and non-smokers there are.
8. Retrieve the Patient_id of patients who have a BMI greaterthan the average BMI
9. Find the patient with the highest HbA1c level and the patient with the lowest HbA1clevel.

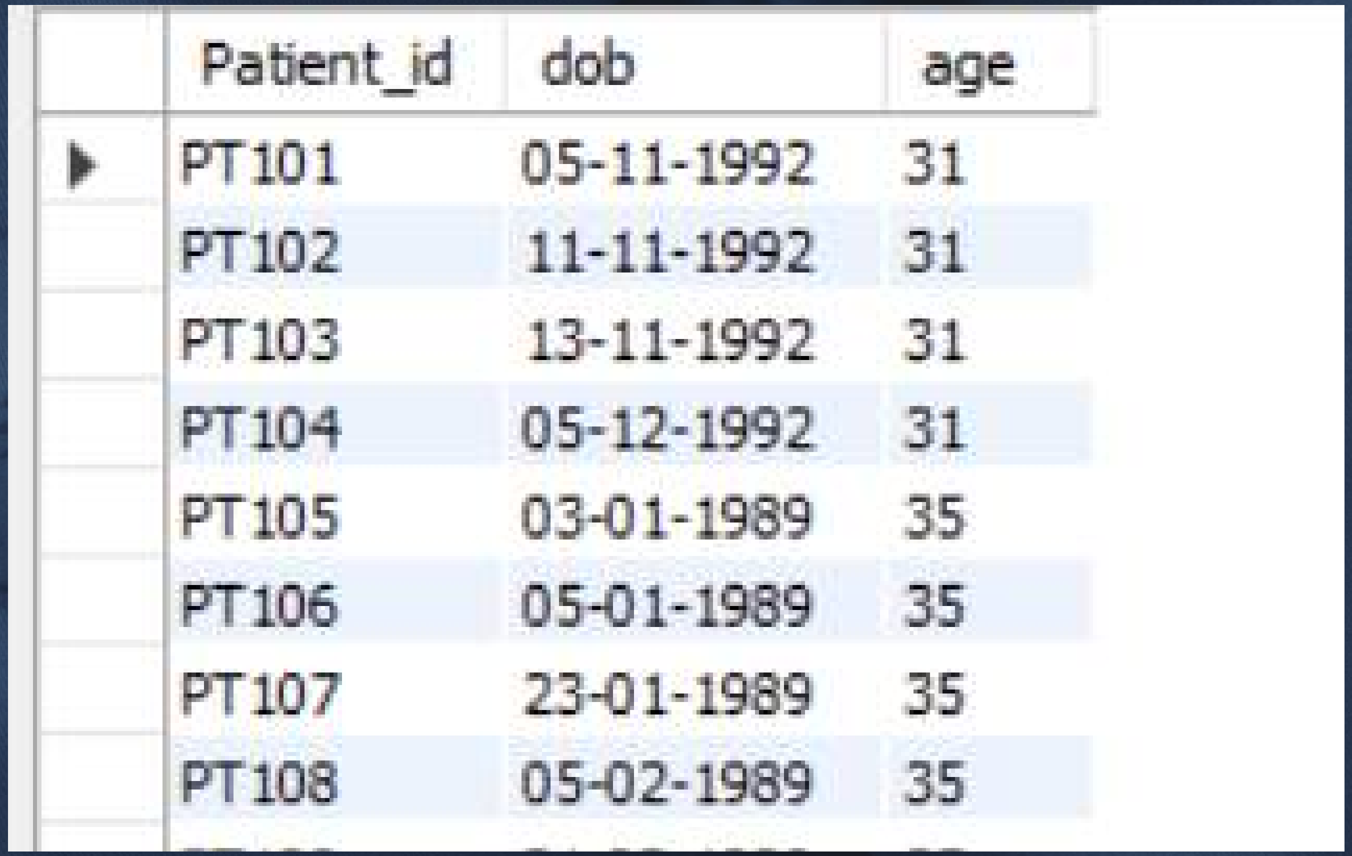
All Questions

10. Calculate the age of patients in years (assuming the current date as of now).
11. Rank patients by blood glucose level within each gender group.
12. Update the smoking history of patients who are olderthan 40 to "Ex-smoker."
13. Insert a new patient into the database with sample data.
14. Delete all patients with heart disease from the database.
15. Find patients who have hypertension but not diabetes using the EXCEPT operator
16. Define a unique constraint on the "patient_id" column to ensure its values are unique.
17. Create a view that displays the Patient_ids, ages, and BMI of patients

Q1. Retrieve the Patient_id and ages of all patients.

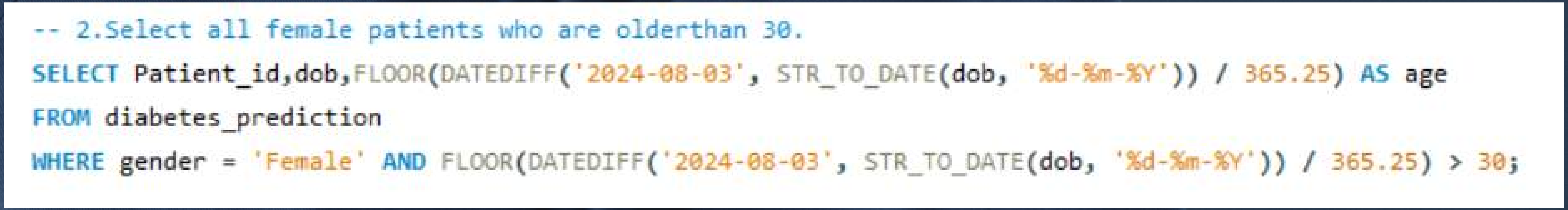


```sql
-- 1. Retrieve the Patient_id and ages of all patients.
SELECT Patient_id,dob,FLOOR(DATEDIFF('2024-08-03', STR_TO_DATE(dob, '%d-%m-%Y')) / 365.25) AS age
FROM diabetes_prediction;
```

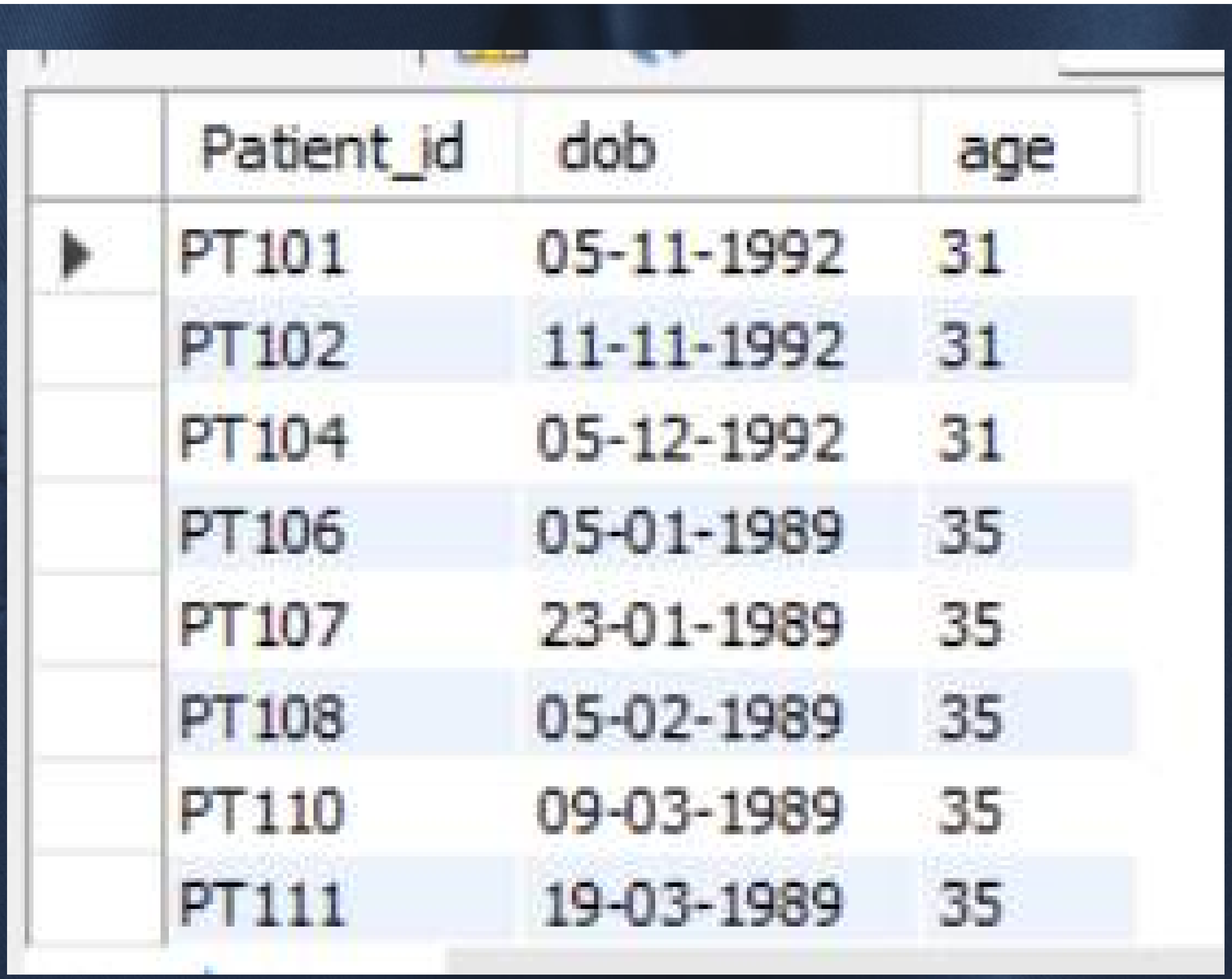


Patient_id	dob	age
PT101	05-11-1992	31
PT102	11-11-1992	31
PT103	13-11-1992	31
PT104	05-12-1992	31
PT105	03-01-1989	35
PT106	05-01-1989	35
PT107	23-01-1989	35
PT108	05-02-1989	35

Q.2 Select all female patients who are olderthan 30.

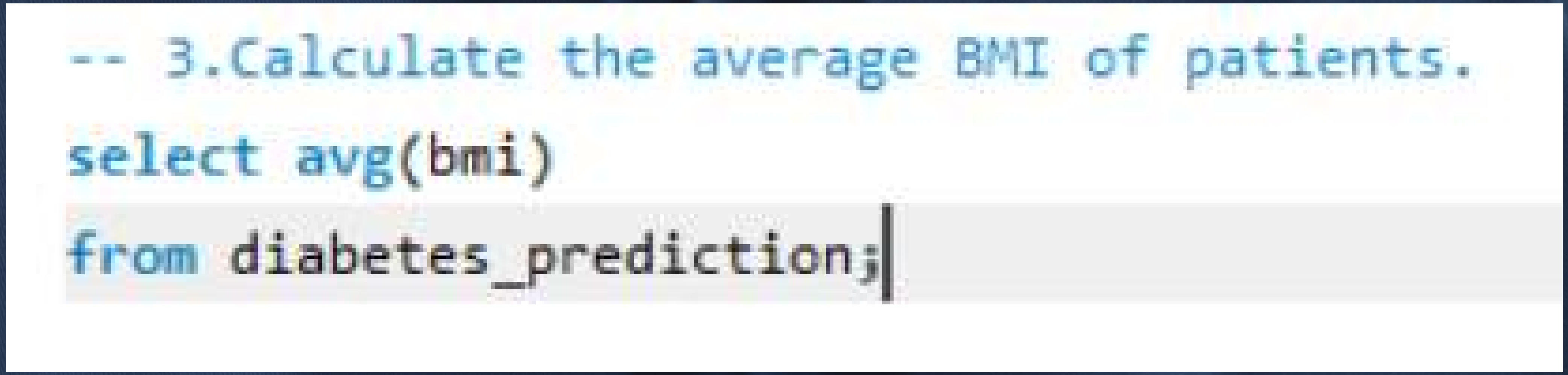


```sql
-- 2.Select all female patients who are olderthan 30.
SELECT Patient_id,dob,FLOOR(DATEDIFF('2024-08-03', STR_TO_DATE(dob, '%d-%m-%Y')) / 365.25) AS age
FROM diabetes_prediction
WHERE gender = 'Female' AND FLOOR(DATEDIFF('2024-08-03', STR_TO_DATE(dob, '%d-%m-%Y')) / 365.25) > 30;
```

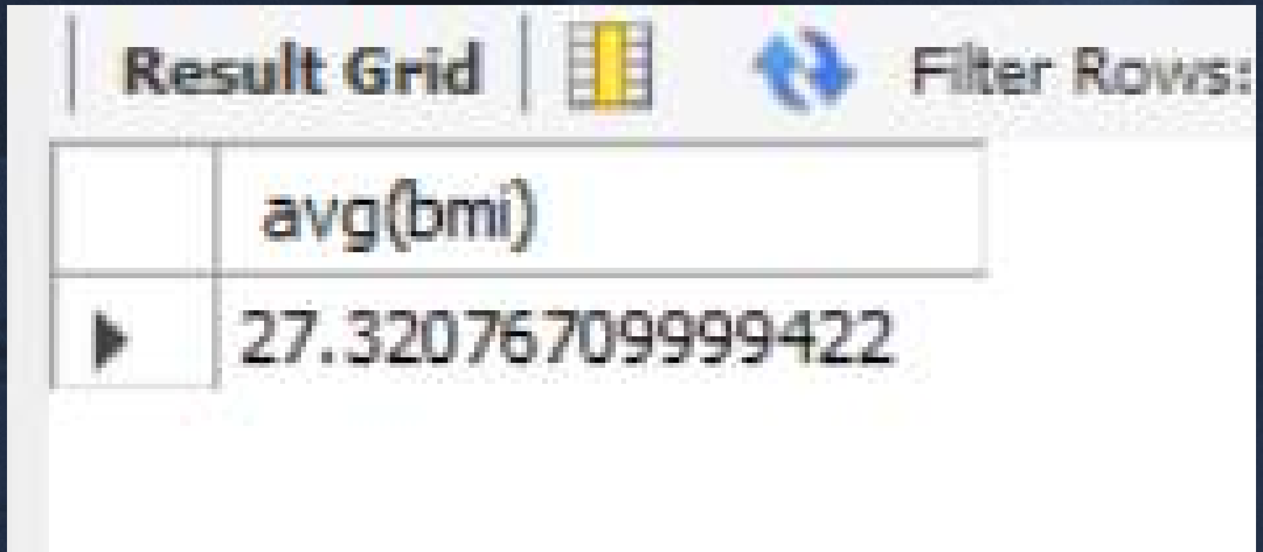


Patient_id	dob	age
PT101	05-11-1992	31
PT102	11-11-1992	31
PT104	05-12-1992	31
PT106	05-01-1989	35
PT107	23-01-1989	35
PT108	05-02-1989	35
PT110	09-03-1989	35
PT111	19-03-1989	35

Q3. Calculate the average BMI of patients.

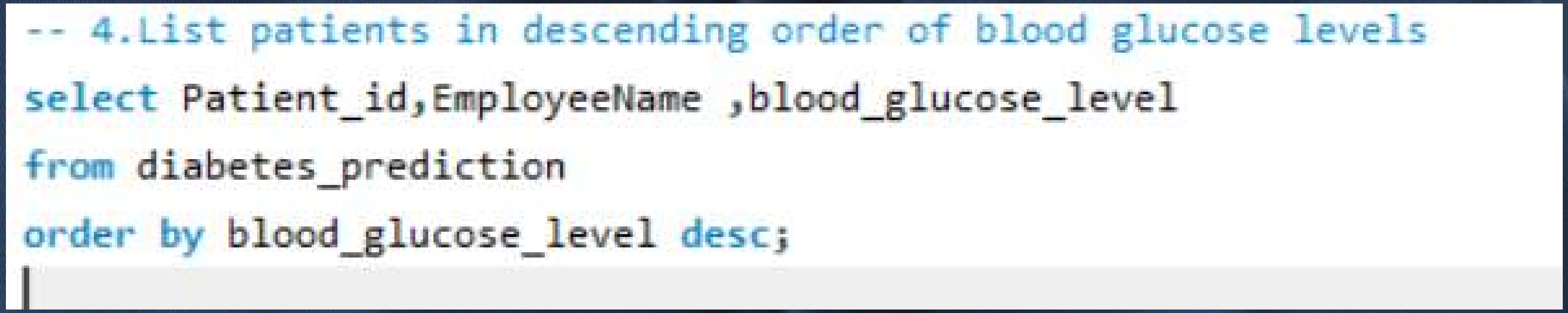


```sql
-- 3.Calculate the average BMI of patients.
select avg(bmi)
from diabetes_prediction;
```

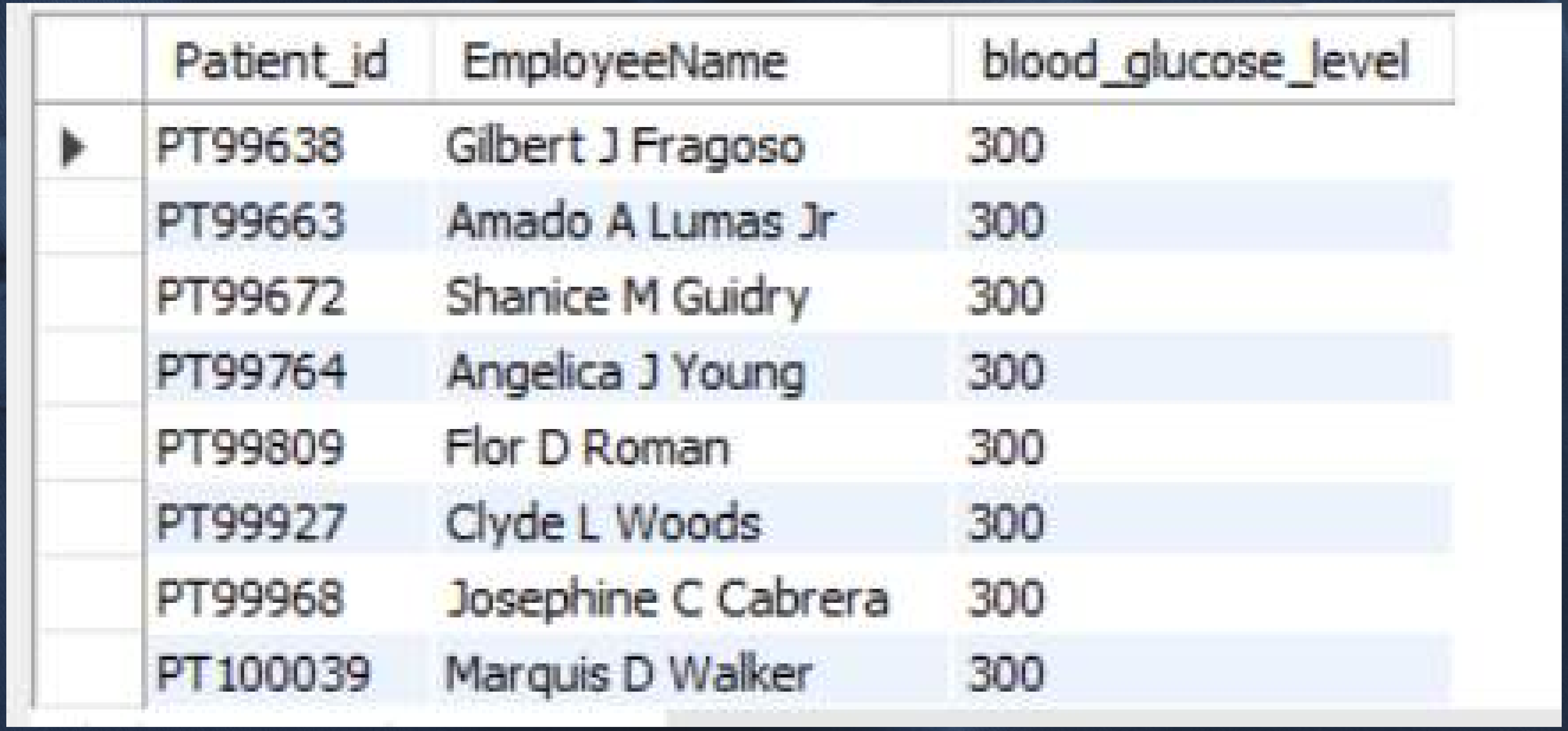


```
avg(bmi)
27.3207670999422
```

Q4. List patients in descending order of blood glucose levels.ges of all patients.

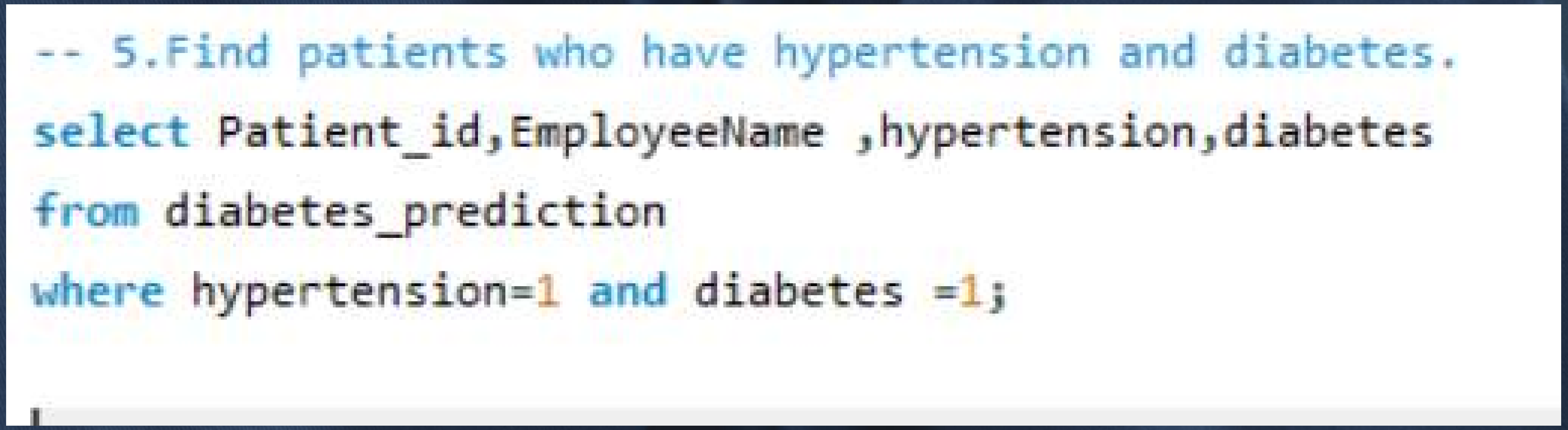


```
-- 4.List patients in descending order of blood glucose levels
select Patient_id,EmployeeName ,blood_glucose_level
from diabetes_prediction
order by blood_glucose_level desc;
```

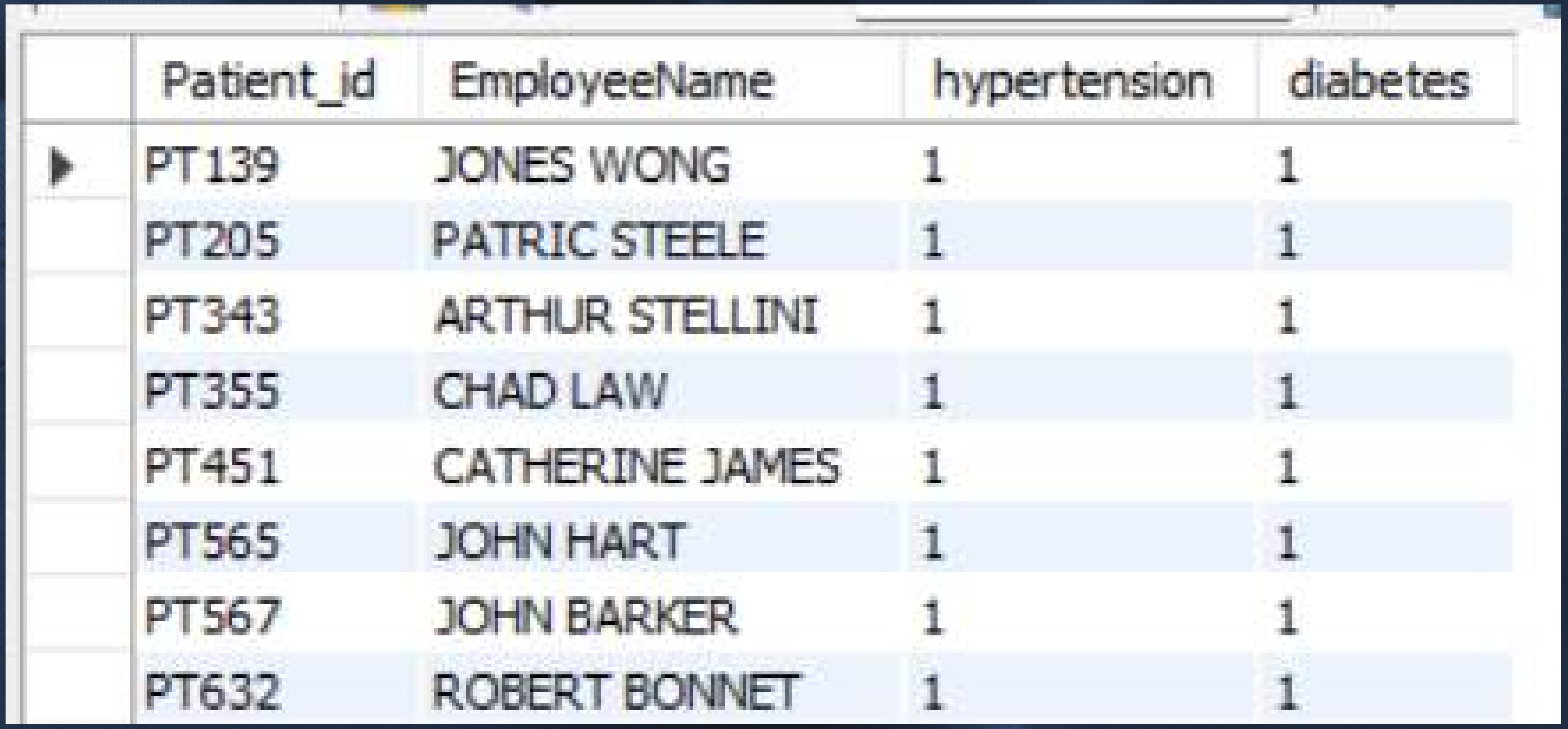


	Patient_id	EmployeeName	blood_glucose_level
▶	PT99638	Gilbert J Fragoso	300
	PT99663	Amado A Lumas Jr	300
	PT99672	Shanice M Guidry	300
	PT99764	Angelica J Young	300
	PT99809	Flor D Roman	300
	PT99927	Clyde L Woods	300
	PT99968	Josephine C Cabrera	300
	PT100039	Marquis D Walker	300

Q5. Find patients who have hypertension and diabetes.

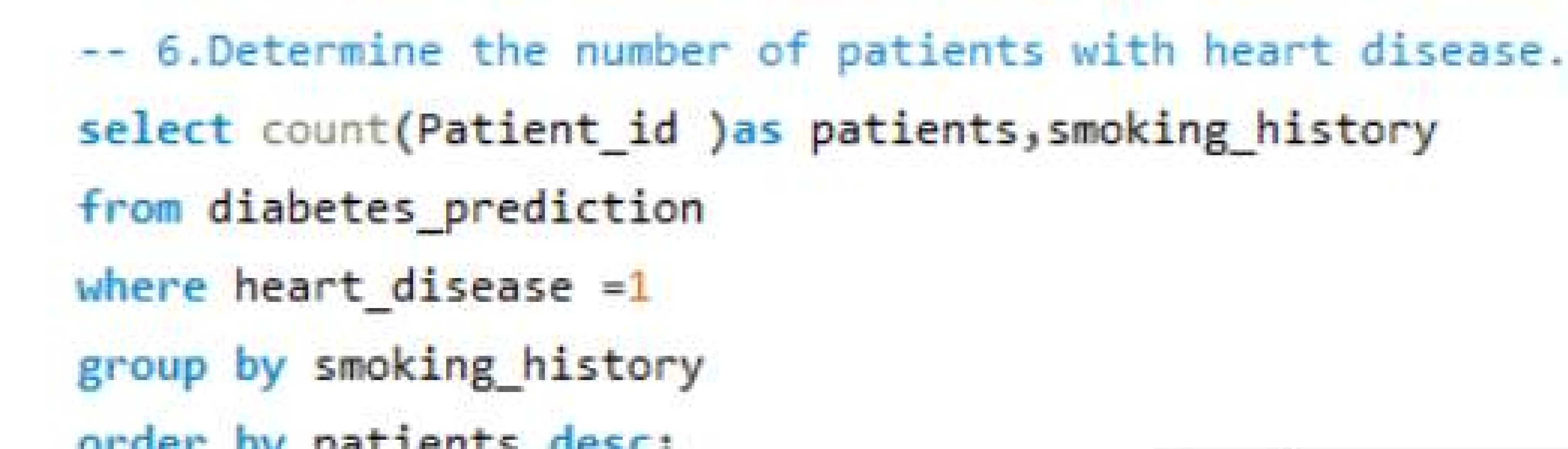


```
-- 5.Find patients who have hypertension and diabetes.
select Patient_id,EmployeeName ,hypertension,diabetes
from diabetes_prediction
where hypertension=1 and diabetes =1;
```

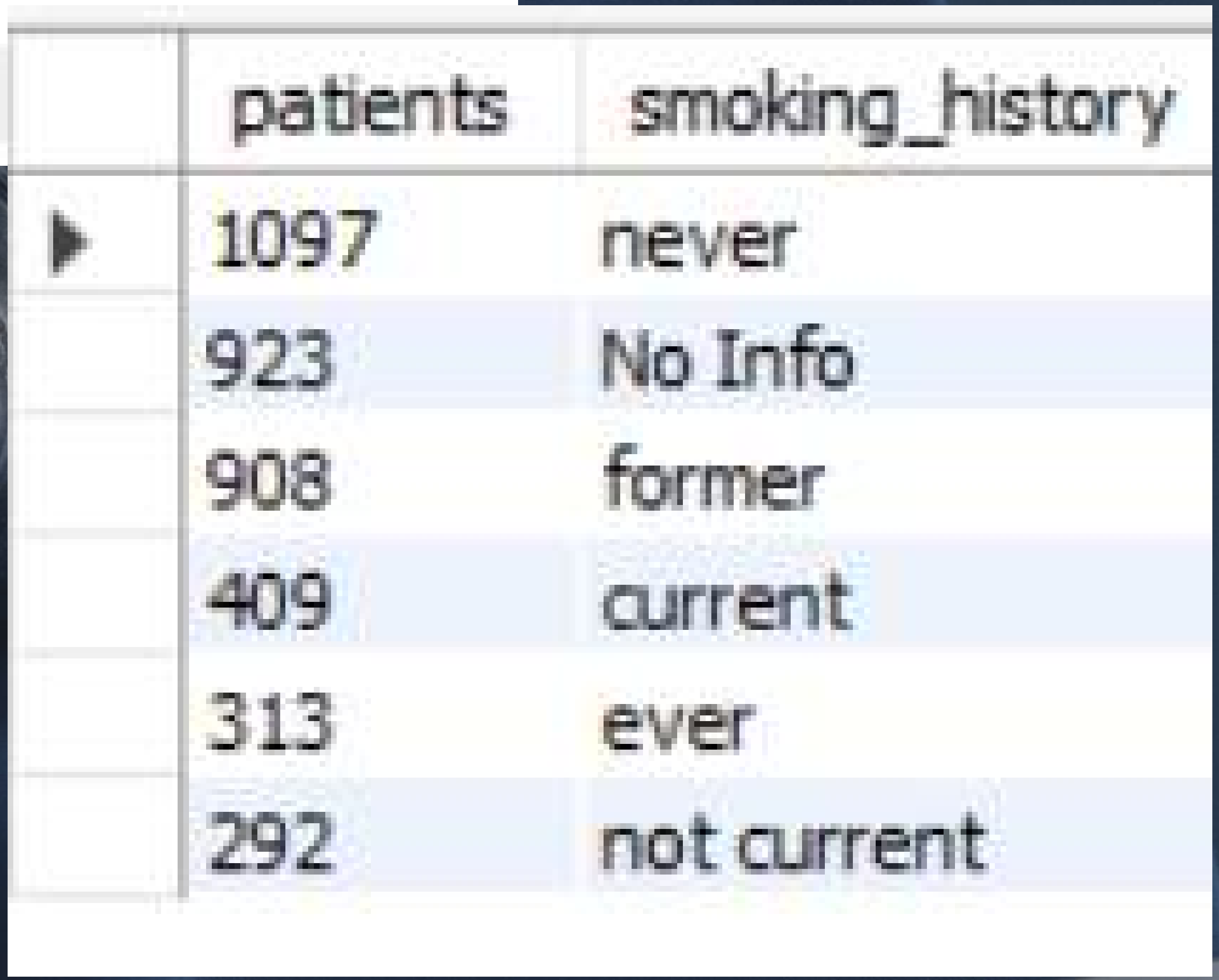


	Patient_id	EmployeeName	hypertension	diabetes
▶	PT139	JONES WONG	1	1
	PT205	PATRIC STEELE	1	1
	PT343	ARTHUR STELLINI	1	1
	PT355	CHAD LAW	1	1
	PT451	CATHERINE JAMES	1	1
	PT565	JOHN HART	1	1
	PT567	JOHN BARKER	1	1
	PT632	ROBERT BONNET	1	1

Q6. Determine the number of patients with heart disease.

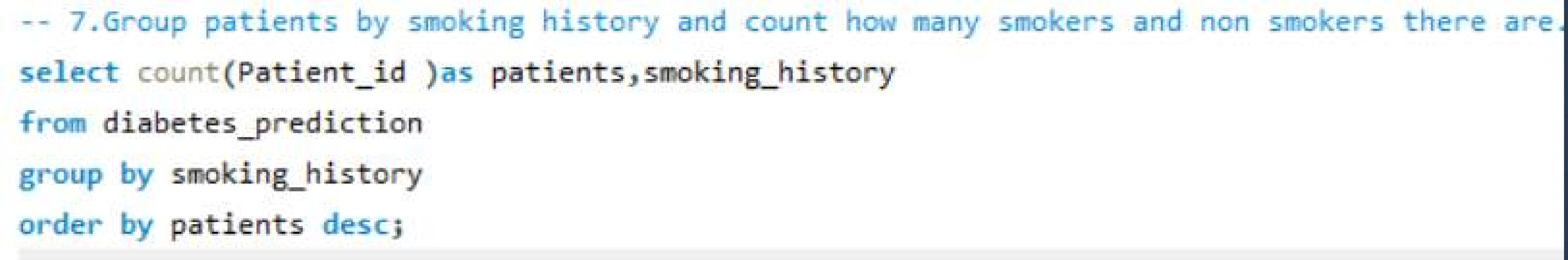


```sql
-- 6.Determine the number of patients with heart disease.
select count(Patient_id )as patients,smoking_history
from diabetes_prediction
where heart_disease =1
group by smoking_history
order by patients desc;
```

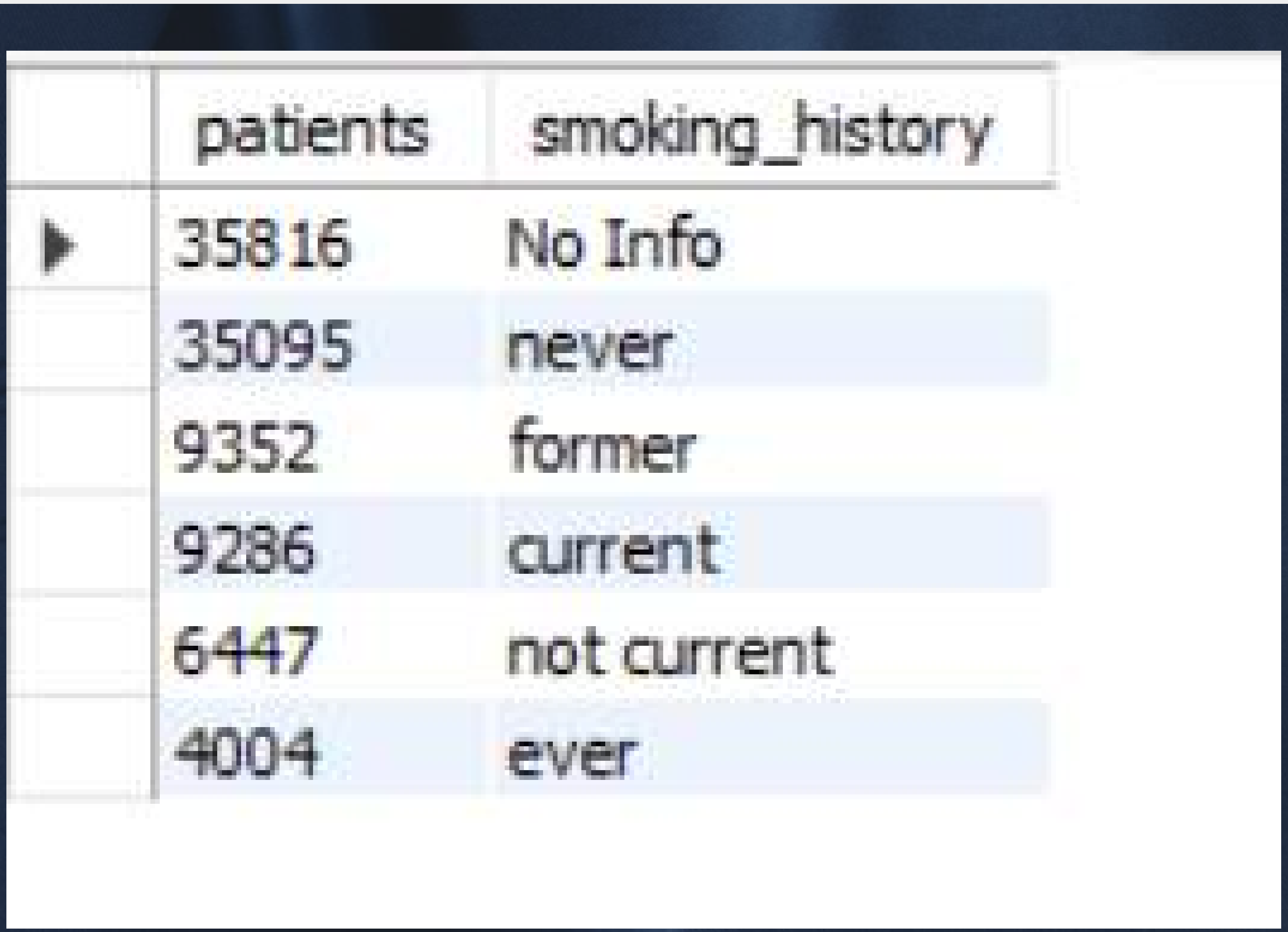


patients	smoking_history
1097	never
923	No Info
908	former
409	current
313	ever
292	not current

Q7. Group patients by smoking history and count how many smokers and non-smokers there are.ts.

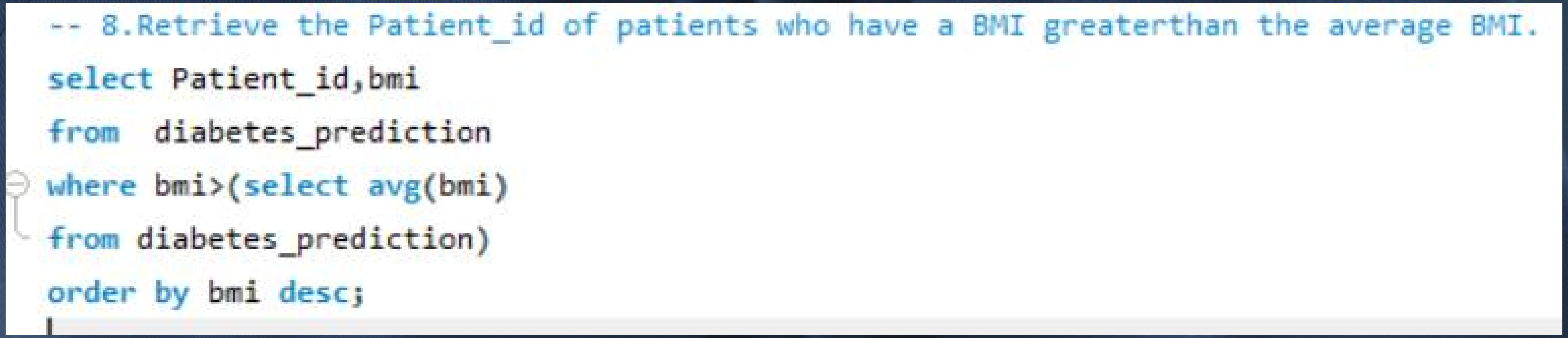


```
-- 7.Group patients by smoking history and count how many smokers and non smokers there are.
select count(Patient_id )as patients,smoking_history
from diabetes_prediction
group by smoking_history
order by patients desc;
```

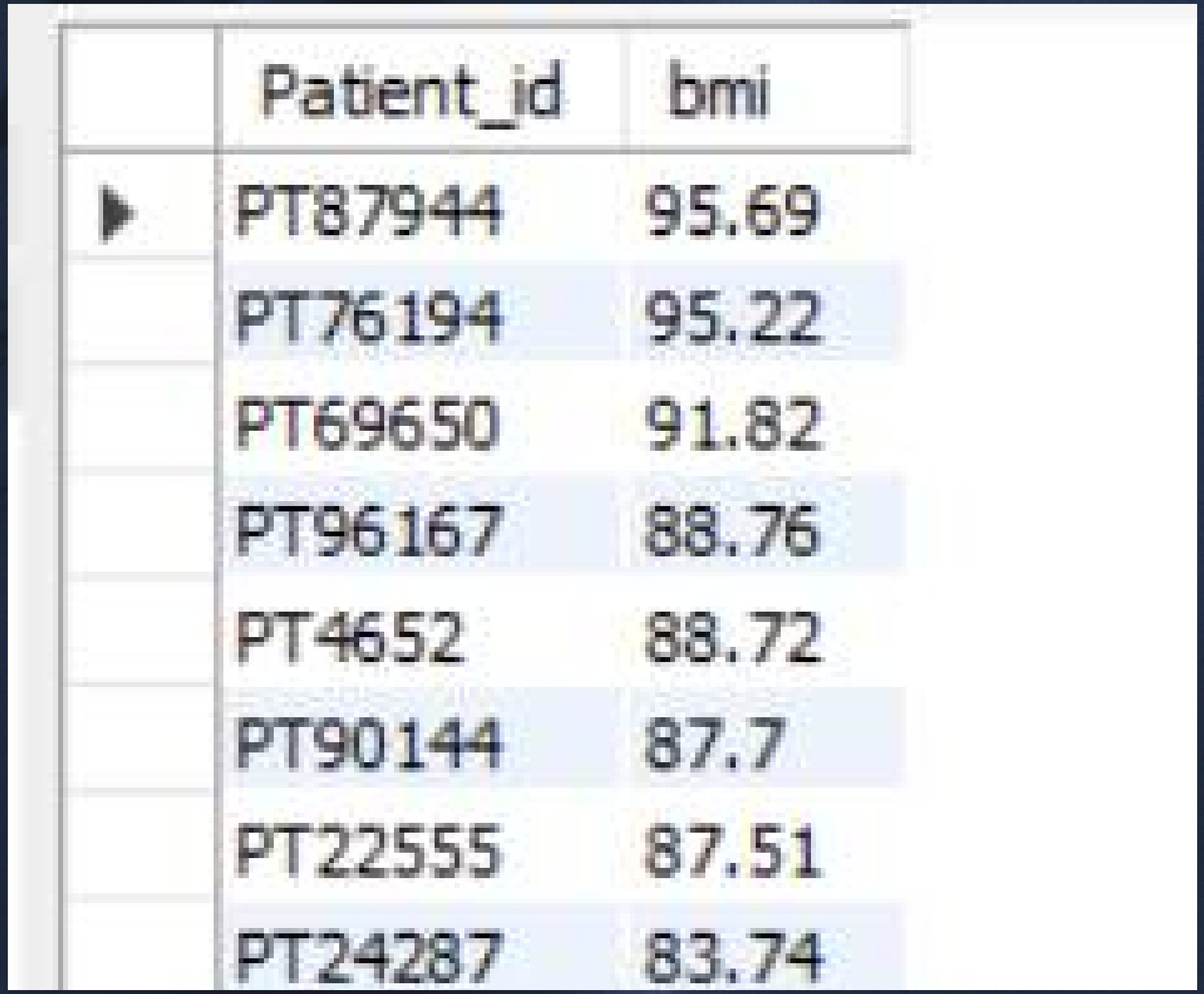


patients	smoking_history
35816	No Info
35095	never
9352	former
9286	current
6447	not current
4004	ever

Q8. Retrieve the Patient_id of patients who have a BMI greaterthan the average BMI

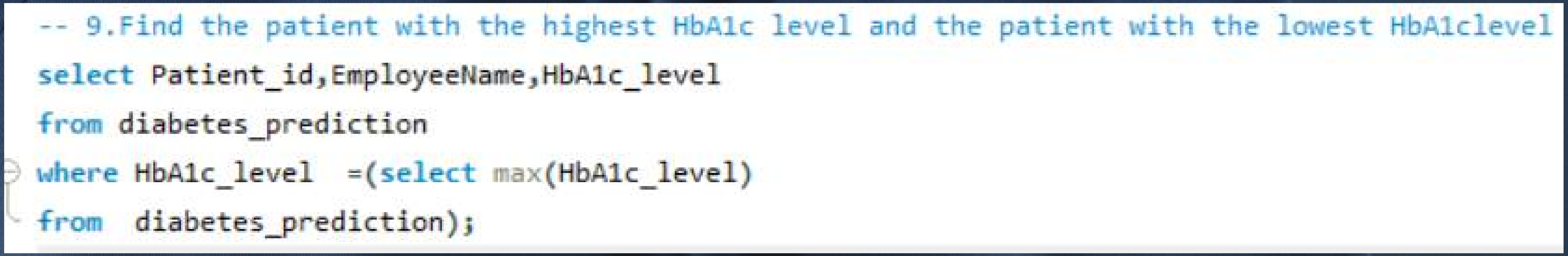


```
-- 8.Retrieve the Patient_id of patients who have a BMI greaterthan the average BMI.
select Patient_id,bmi
from  diabetes_prediction
where bmi>(select avg(bmi)
from diabetes_prediction)
order by bmi desc;
```

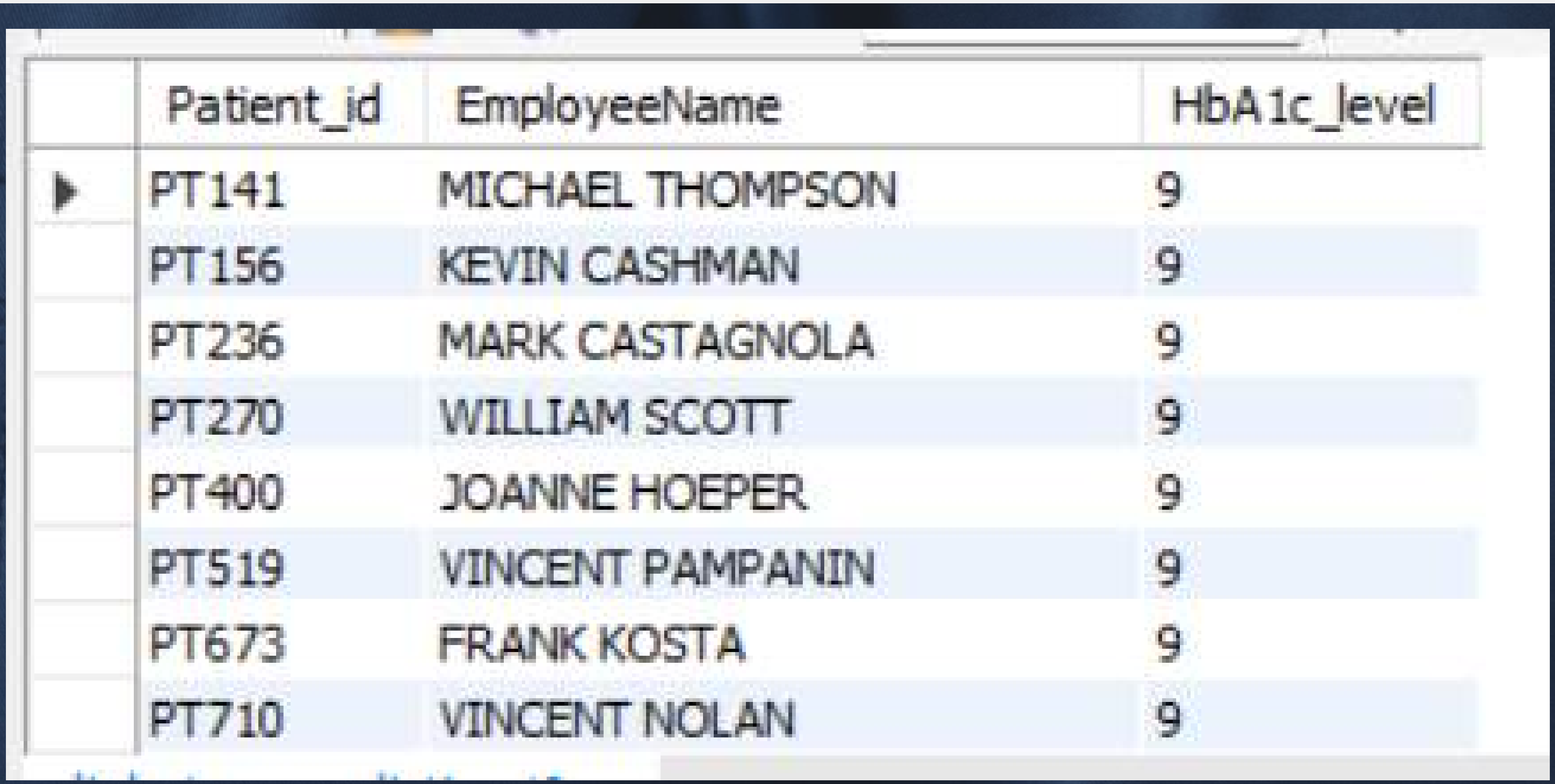


Patient_id	bmi
PT87944	95.69
PT76194	95.22
PT69650	91.82
PT96167	88.76
PT4652	88.72
PT90144	87.7
PT22555	87.51
PT24287	83.74

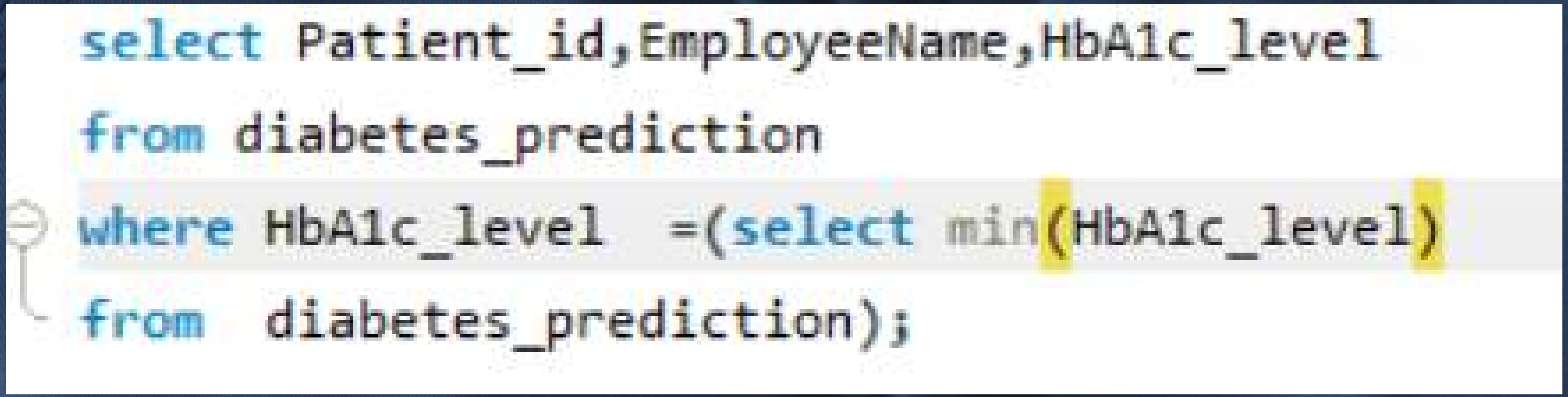
Q9. Find the patient with the highest HbA1c level and the patient with the lowest HbA1clevel.



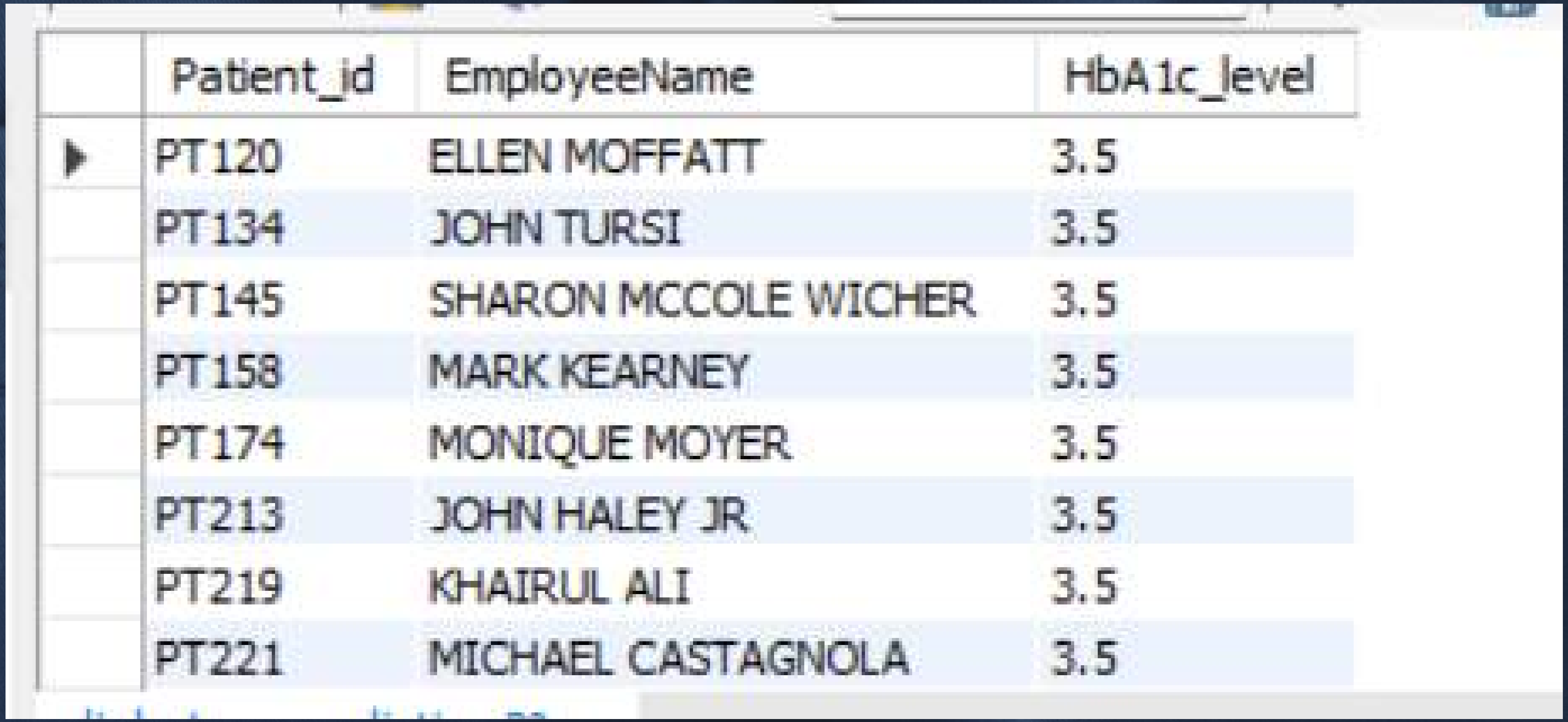
```
-- 9.Find the patient with the highest HbA1c level and the patient with the lowest HbA1clevel
select Patient_id,EmployeeName,HbA1c_level
from diabetes_prediction
where HbA1c_level  =(select max(HbA1c_level)
from  diabetes_prediction);
```



Patient_id	EmployeeName	HbA1c_level
PT141	MICHAEL THOMPSON	9
PT156	KEVIN CASHMAN	9
PT236	MARK CASTAGNOLA	9
PT270	WILLIAM SCOTT	9
PT400	JOANNE HOEPER	9
PT519	VINCENT PAMPANIN	9
PT673	FRANK KOSTA	9
PT710	VINCENT NOLAN	9

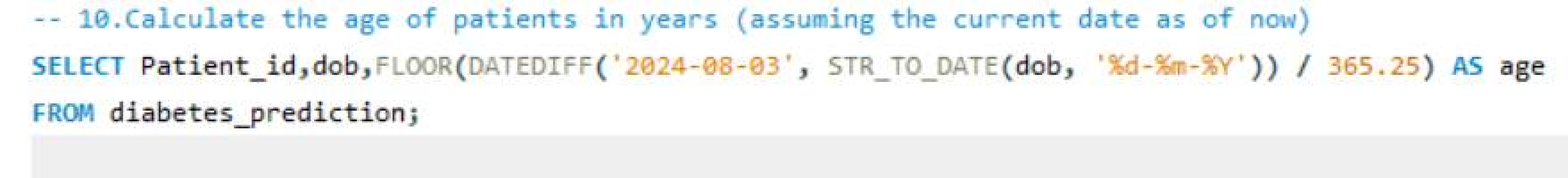


```
select Patient_id,EmployeeName,HbA1c_level
from diabetes_prediction
where HbA1c_level  =(select min(HbA1c_level)
from  diabetes_prediction);
```

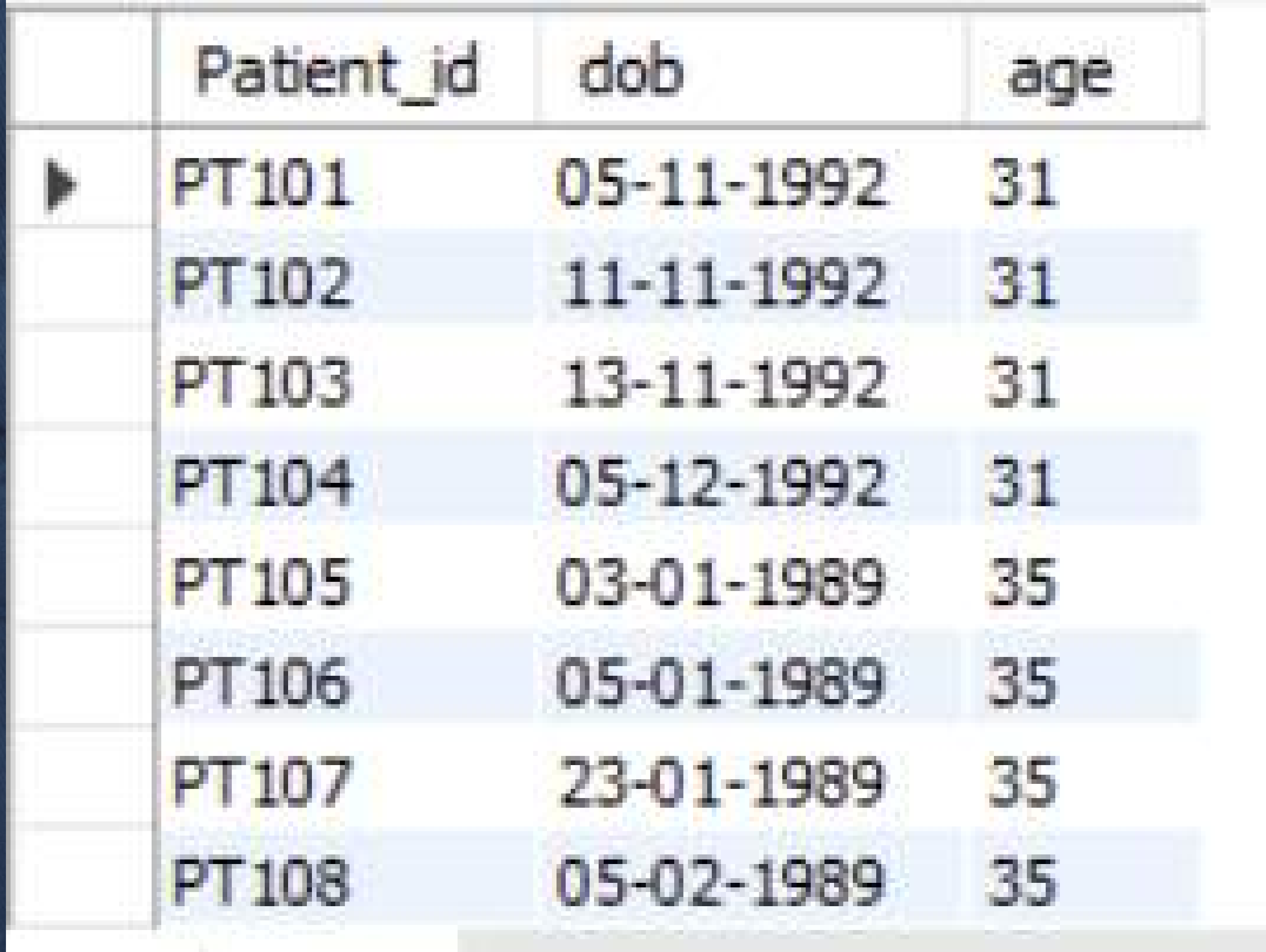


Patient_id	EmployeeName	HbA1c_level
PT120	ELLEN MOFFATT	3.5
PT134	JOHN TURSI	3.5
PT145	SHARON MCCOLE WICHER	3.5
PT158	MARK KEARNEY	3.5
PT174	MONIQUE MOYER	3.5
PT213	JOHN HALEY JR	3.5
PT219	KHAIRUL ALI	3.5
PT221	MICHAEL CASTAGNOLA	3.5

Q10. Calculate the age of patients in years (assuming the current date as of now).

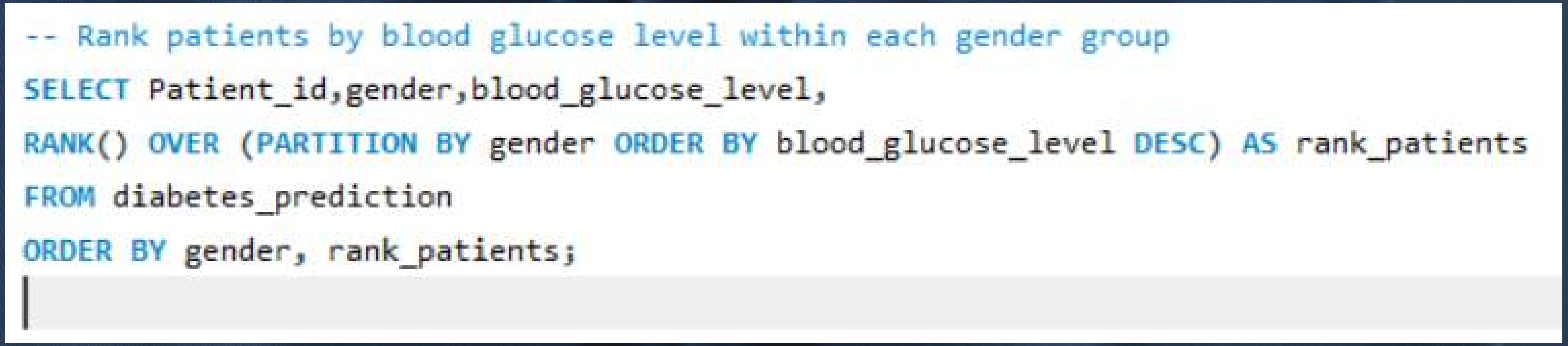


```sql
-- 10.Calculate the age of patients in years (assuming the current date as of now)
SELECT Patient_id,dob,FLOOR(DATEDIFF('2024-08-03', STR_TO_DATE(dob, '%d-%m-%Y')) / 365.25) AS age
FROM diabetes_prediction;
```

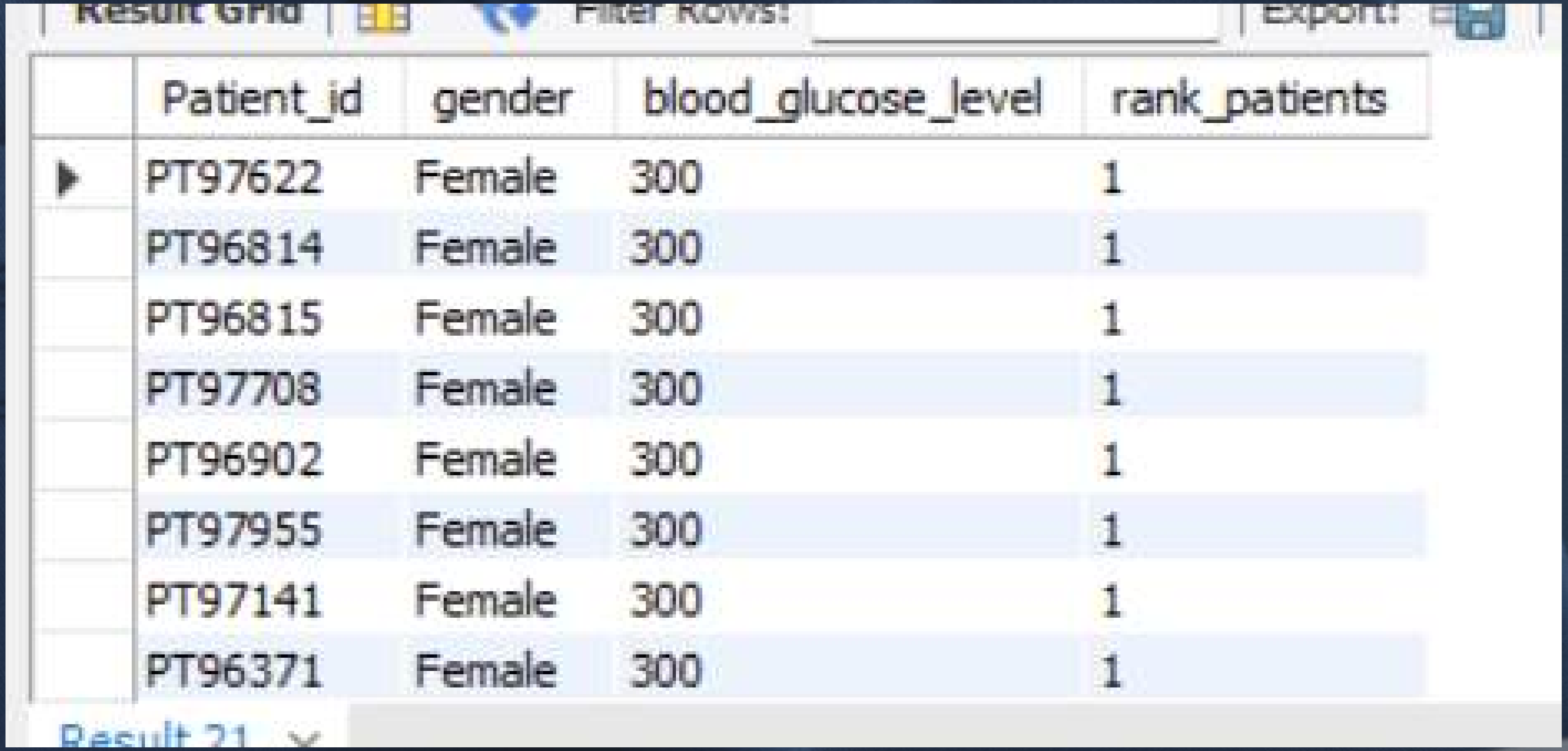


Patient_id	dob	age
PT101	05-11-1992	31
PT102	11-11-1992	31
PT103	13-11-1992	31
PT104	05-12-1992	31
PT105	03-01-1989	35
PT106	05-01-1989	35
PT107	23-01-1989	35
PT108	05-02-1989	35

Q11. Rank patients by blood glucose level within each gender group.

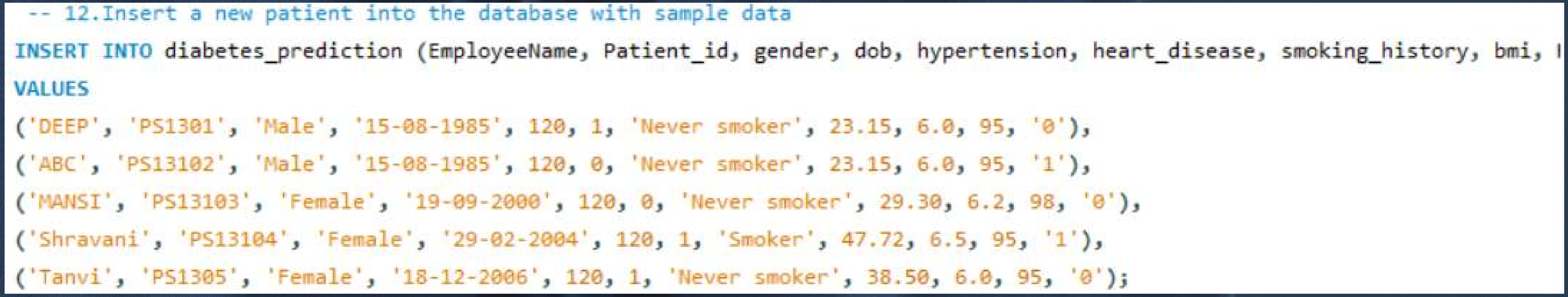


```
-- Rank patients by blood glucose level within each gender group
SELECT Patient_id,gender,blood_glucose_level,
RANK() OVER (PARTITION BY gender ORDER BY blood_glucose_level DESC) AS rank_patients
FROM diabetes_prediction
ORDER BY gender, rank_patients;
```



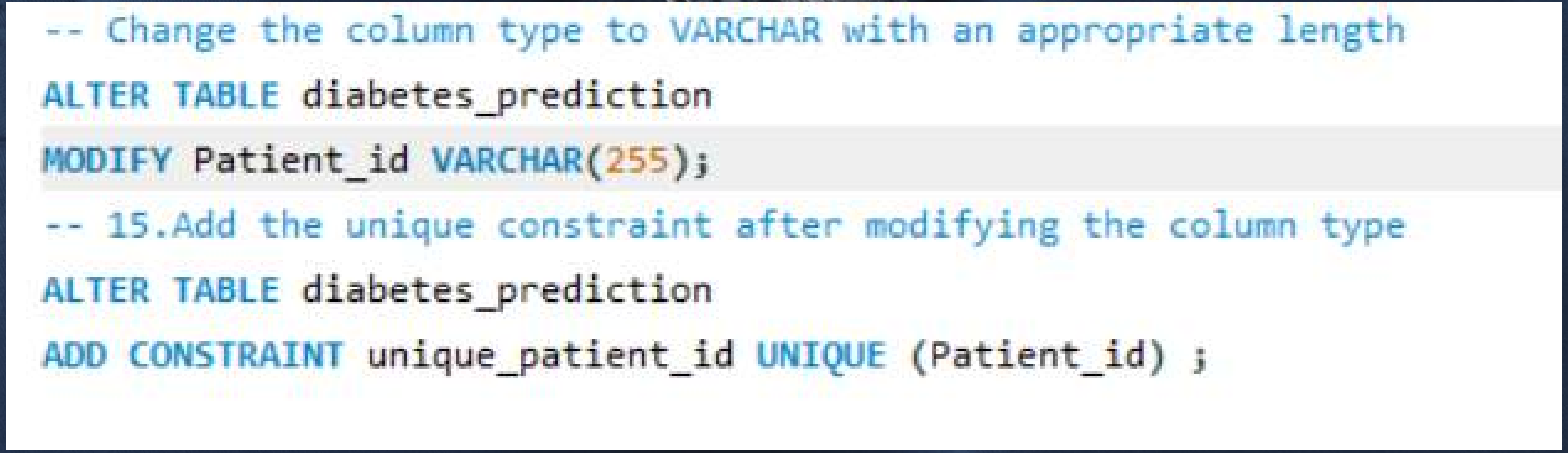
Patient_id	gender	blood_glucose_level	rank_patients
PT97622	Female	300	1
PT96814	Female	300	1
PT96815	Female	300	1
PT97708	Female	300	1
PT96902	Female	300	1
PT97955	Female	300	1
PT97141	Female	300	1
PT96371	Female	300	1

Q13. Insert a new patient into the database with sample data.



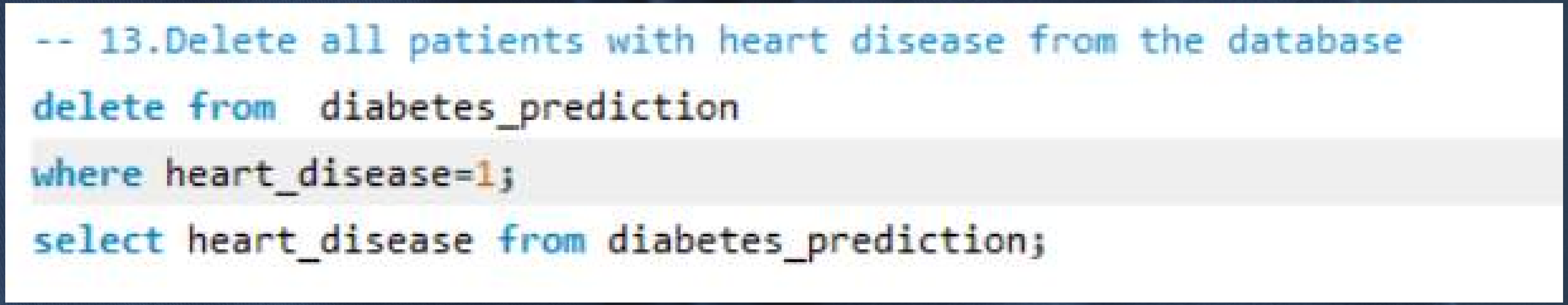
```
 -- 12.Insert a new patient into the database with sample data
INSERT INTO diabetes_prediction (EmployeeName, Patient_id, gender, dob, hypertension, heart_disease, smoking_history, bmi,
VALUES
('DEEP', 'PS1301', 'Male', '15-08-1985', 120, 1, 'Never smoker', 23.15, 6.0, 95, '0'),
('ABC', 'PS13102', 'Male', '15-08-1985', 120, 0, 'Never smoker', 23.15, 6.0, 95, '1'),
('MANSI', 'PS13103', 'Female', '19-09-2000', 120, 0, 'Never smoker', 29.30, 6.2, 98, '0'),
('Shravani', 'PS13104', 'Female', '29-02-2004', 120, 1, 'Smoker', 47.72, 6.5, 95, '1'),
('Tanvi', 'PS1305', 'Female', '18-12-2006', 120, 1, 'Never smoker', 38.50, 6.0, 95, '0');
```

Q16. Define a unique constraint on the "patient_id" column to ensure its values are unique

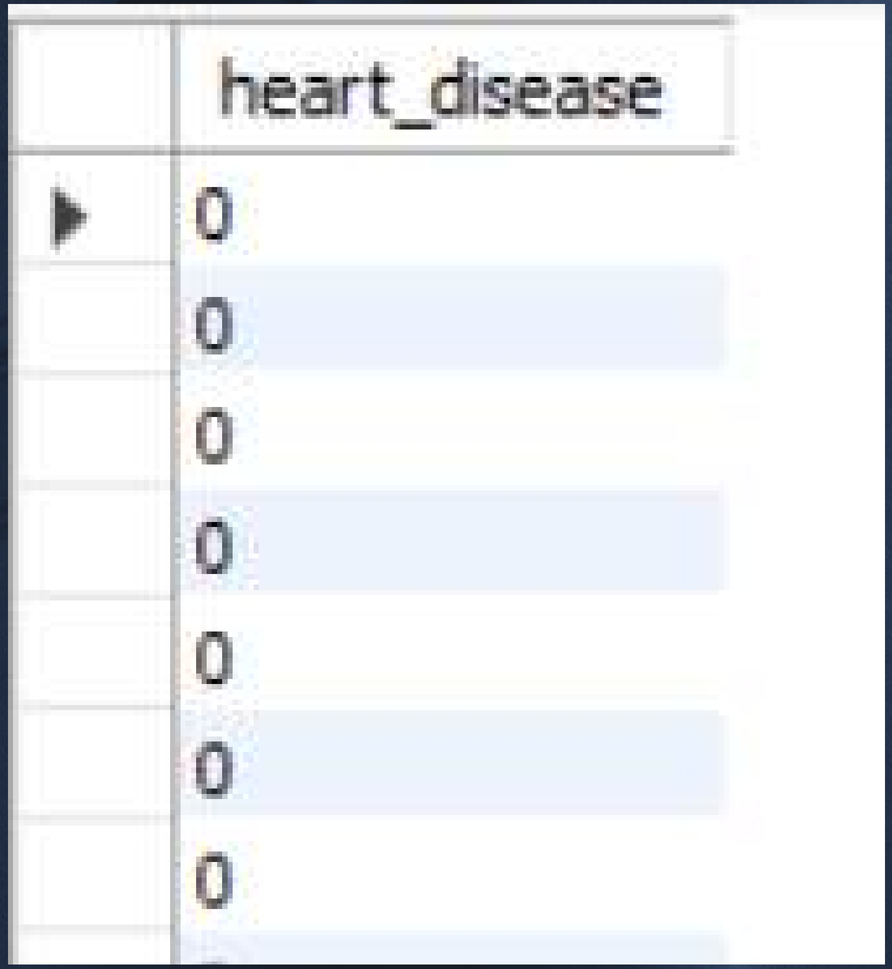


```
-- Change the column type to VARCHAR with an appropriate length
ALTER TABLE diabetes_prediction
MODIFY Patient_id VARCHAR(255);
-- 15.Add the unique constraint after modifying the column type
ALTER TABLE diabetes_prediction
ADD CONSTRAINT unique_patient_id UNIQUE (Patient_id) ;
```

Q14. Delete all patients with heart disease from the database

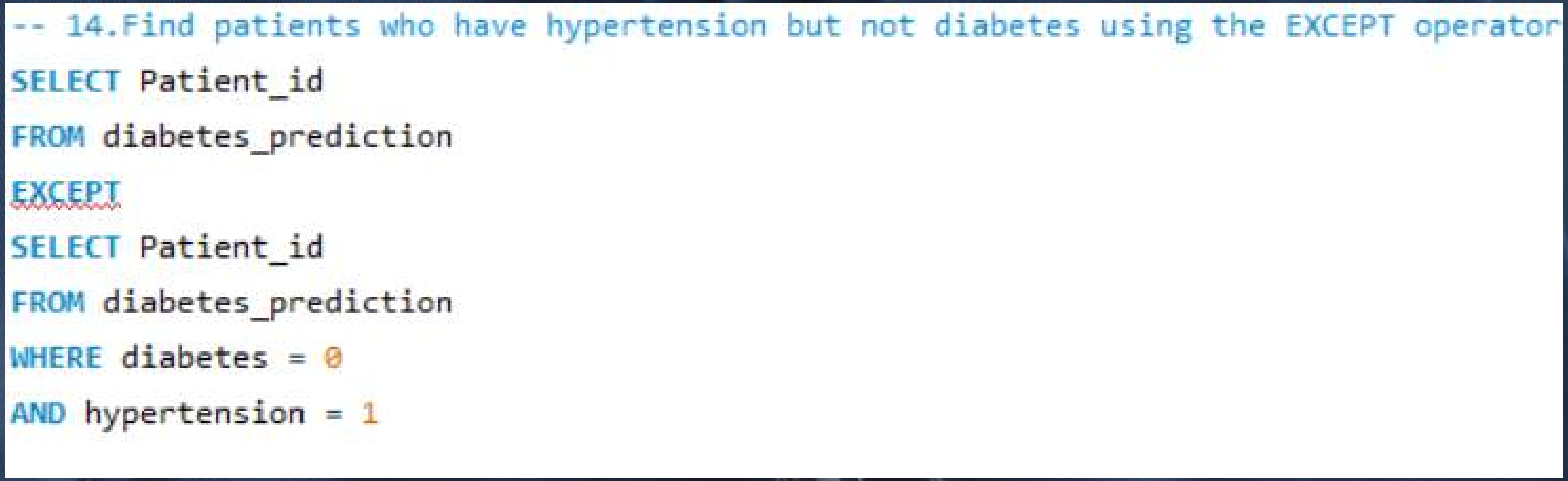


```sql
-- 13.Delete all patients with heart disease from the database
delete from  diabetes_prediction
where heart_disease=1;
select heart_disease from diabetes_prediction;
```

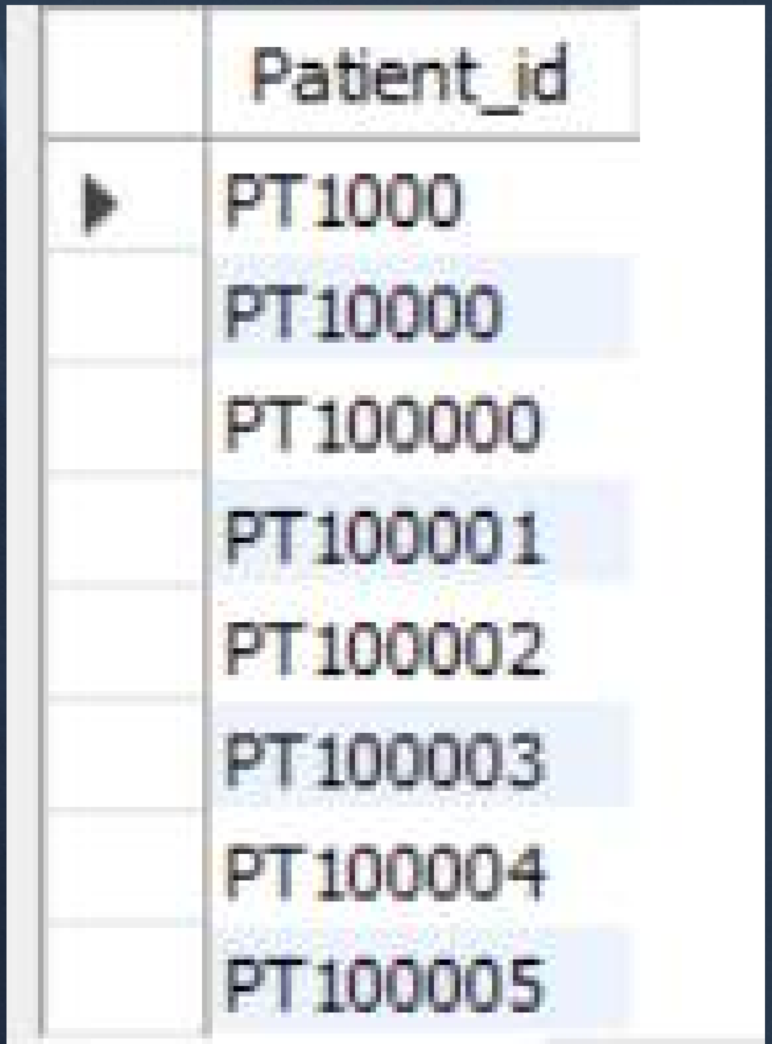


```
   heart_disease
▶  0
   0
   0
   0
   0
   0
   0
```

Q15 Find patients who have hypertension but not diabetes using the EXCEPT operator

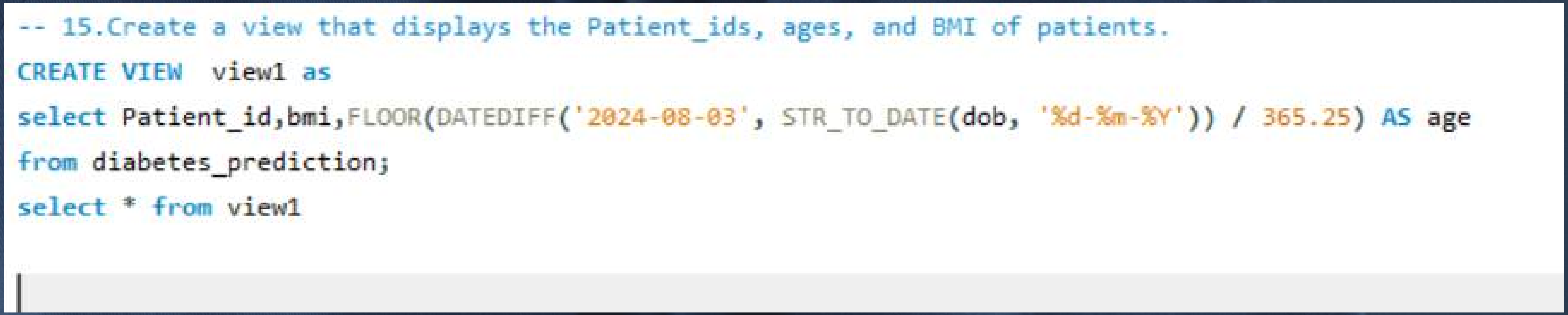


```
-- 14.Find patients who have hypertension but not diabetes using the EXCEPT operator
SELECT Patient_id
FROM diabetes_prediction
EXCEPT
SELECT Patient_id
FROM diabetes_prediction
WHERE diabetes = 0
AND hypertension = 1
```

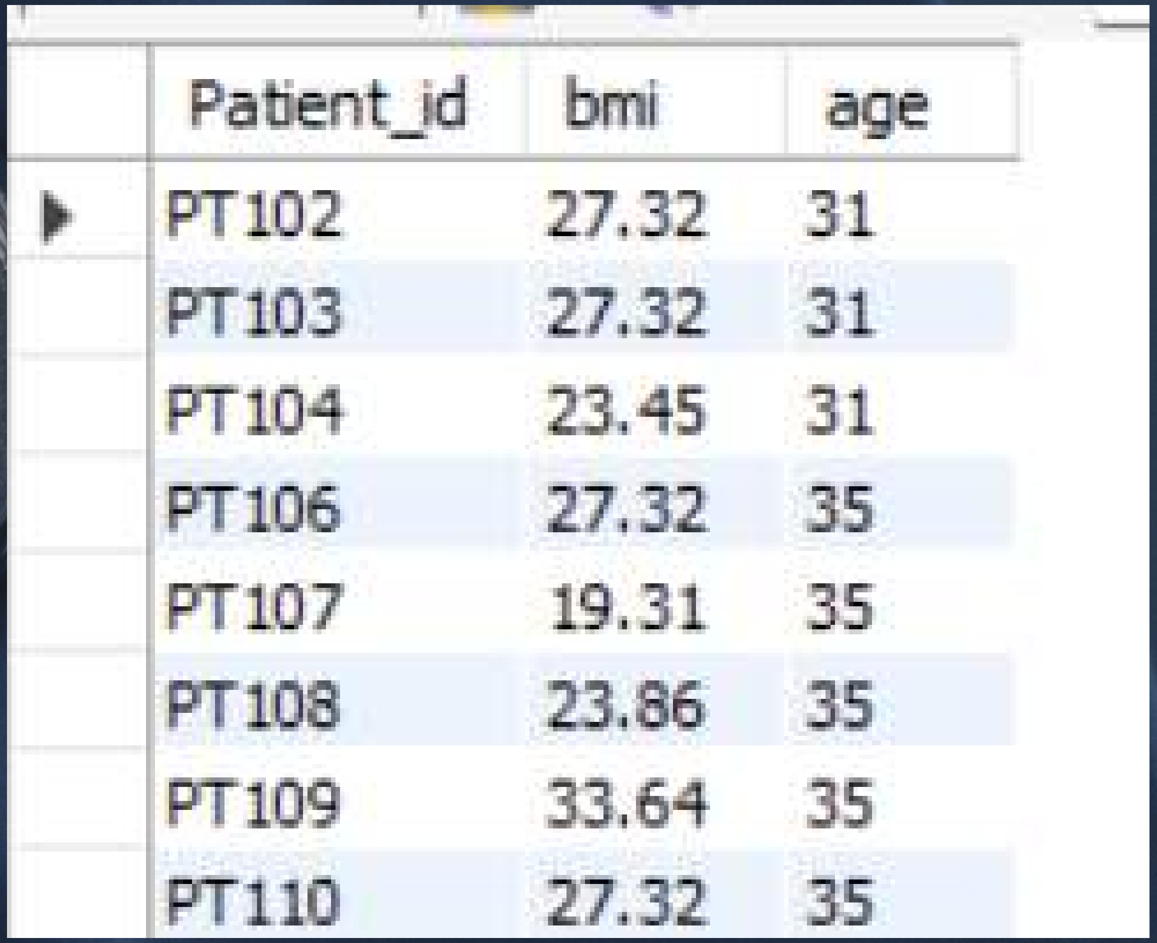


Patient_id
PT1000
PT10000
PT100000
PT100001
PT100002
PT100003
PT100004
PT100005

Q17. Create a view that displays the Patient_ids, ages, and BMI of patient

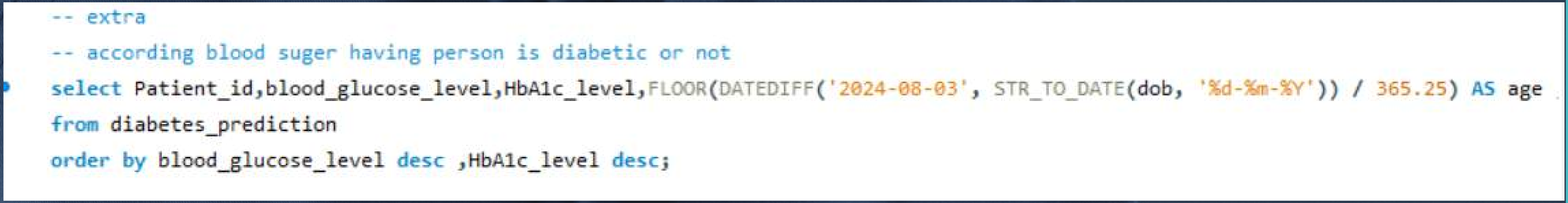


```
-- 15.Create a view that displays the Patient_ids, ages, and BMI of patients.
CREATE VIEW  view1 as
select Patient_id,bmi,FLOOR(DATEDIFF('2024-08-03', STR_TO_DATE(dob, '%d-%m-%Y')) / 365.25) AS age
from diabetes_prediction;
select * from view1
```

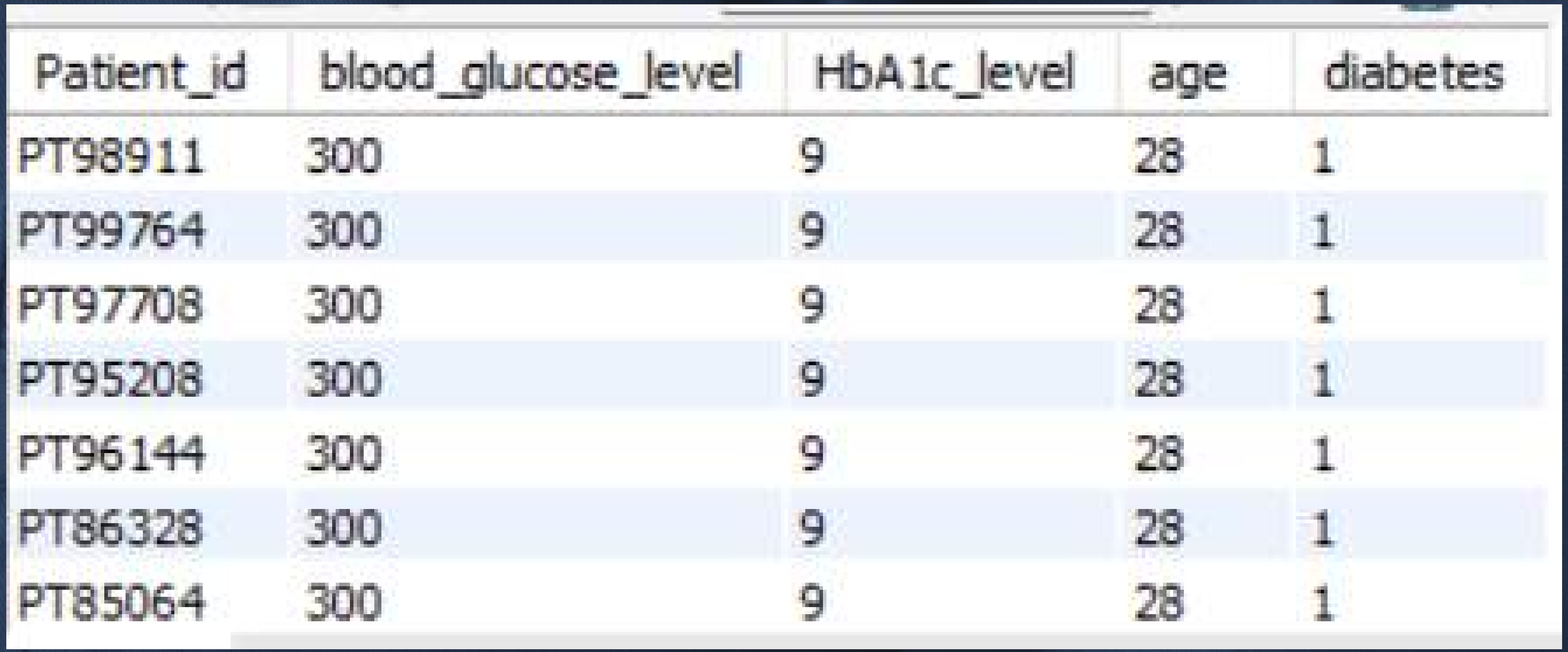


	Patient_id	bmi	age
▶	PT102	27.32	31
	PT103	27.32	31
	PT104	23.45	31
	PT106	27.32	35
	PT107	19.31	35
	PT108	23.86	35
	PT109	33.64	35
	PT110	27.32	35

Q16. Finding out person is diabetic or not .



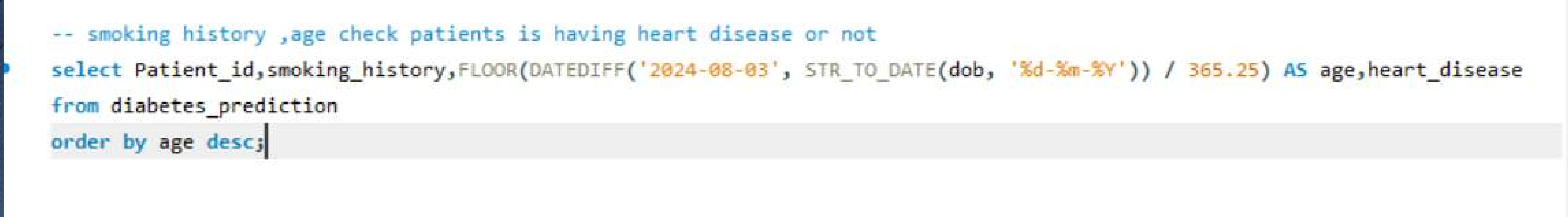
```
-- extra
-- according blood suger having person is diabetic or not
select Patient_id,blood_glucose_level,HbA1c_level,FLOOR(DATEDIFF('2024-08-03', STR_TO_DATE(dob, '%d-%m-%Y')) / 365.25) AS age
from diabetes_prediction
order by blood_glucose_level desc ,HbA1c_level desc;
```



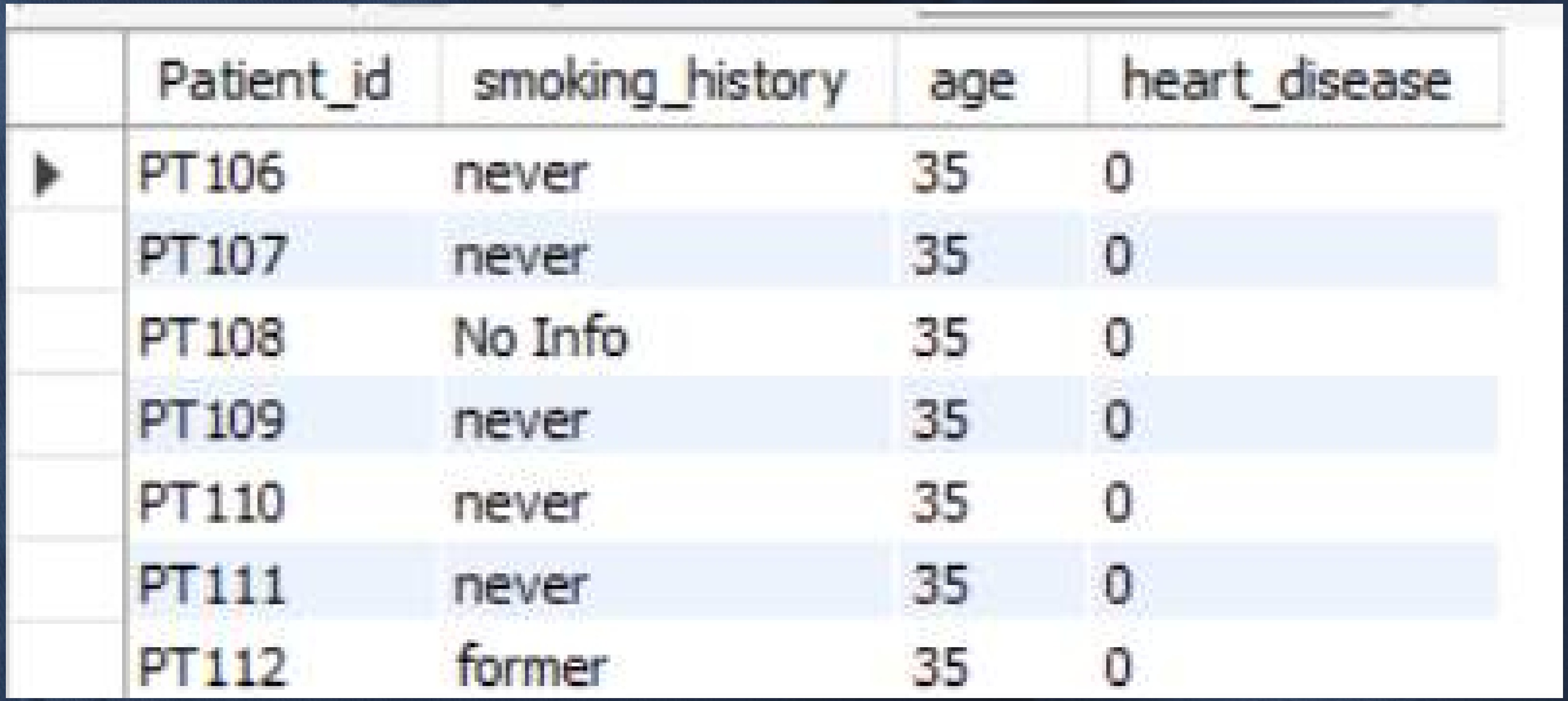
Patient_id	blood_glucose_level	HbA1c_level	age	diabetes
PT98911	300	9	28	1
PT99764	300	9	28	1
PT97708	300	9	28	1
PT95208	300	9	28	1
PT96144	300	9	28	1
PT86328	300	9	28	1
PT85064	300	9	28	1

Patients with the highest blood glucose levels and HbA1c levels are likely to have diabetes. This data shows that, at age 28, these levels are most commonly observed

Q17 Finding patients is having heart_disease accoding thier smoking_history



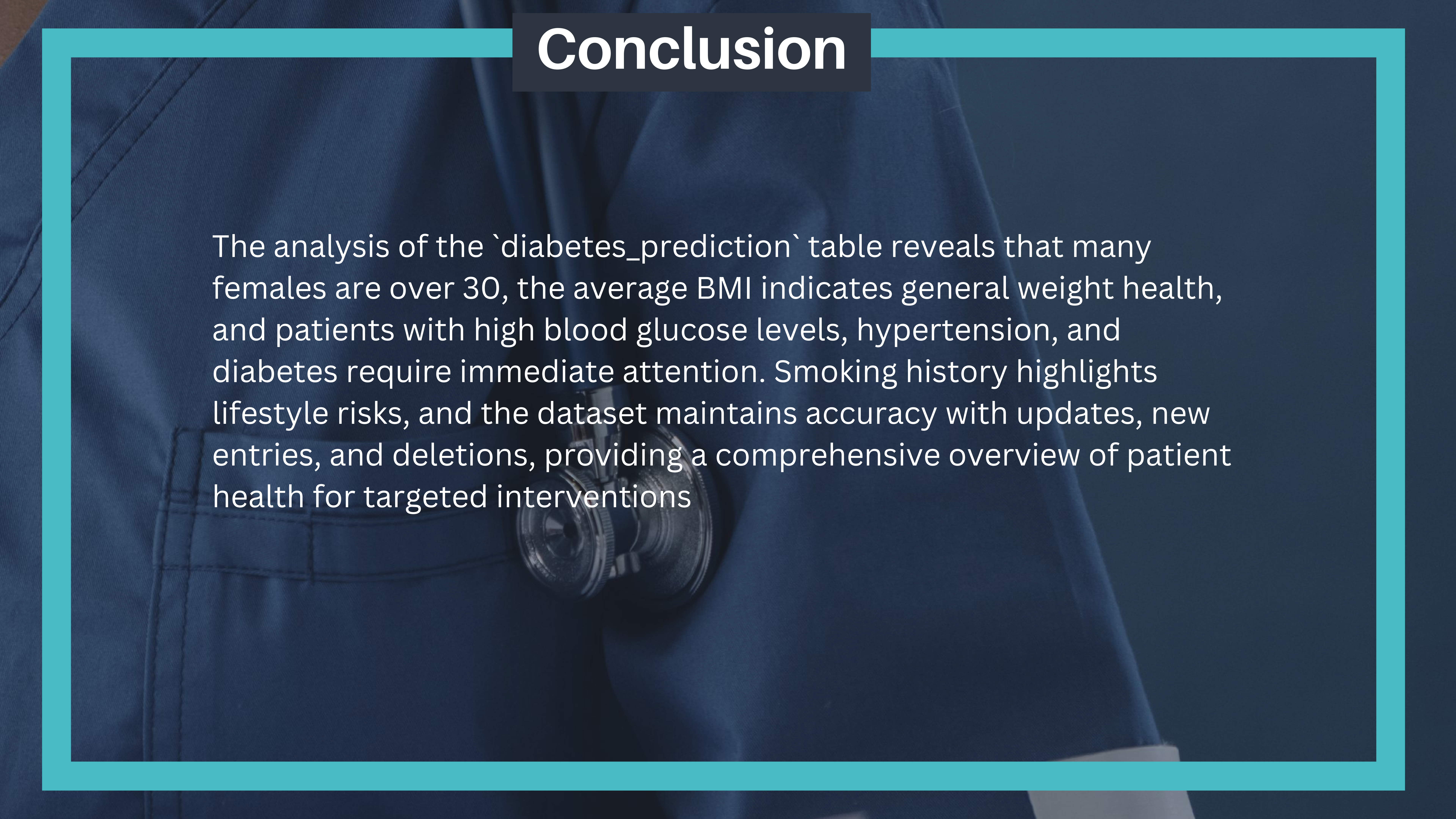
```sql
-- smoking history ,age check patients is having heart disease or not
select Patient_id,smoking_history,FLOOR(DATEDIFF('2024-08-03', STR_TO_DATE(dob, '%d-%m-%Y')) / 365.25) AS age,heart_disease
from diabetes_prediction
order by age desc;
```



Patient_id	smoking_history	age	heart_disease
PT106	never	35	0
PT107	never	35	0
PT108	No Info	35	0
PT109	never	35	0
PT110	never	35	0
PT111	never	35	0
PT112	former	35	0

The most common age for such patients is 35, and if the person currently smokes, this indicates a higher likelihood of heart disease.

Conclusion



The analysis of the `diabetes_prediction` table reveals that many females are over 30, the average BMI indicates general weight health, and patients with high blood glucose levels, hypertension, and diabetes require immediate attention. Smoking history highlights lifestyle risks, and the dataset maintains accuracy with updates, new entries, and deletions, providing a comprehensive overview of patient health for targeted interventions