PSYLIQ

# DIABETES PREDICTION DATASET

BY
MRINAL DESHMUKH

# Basic Introduction

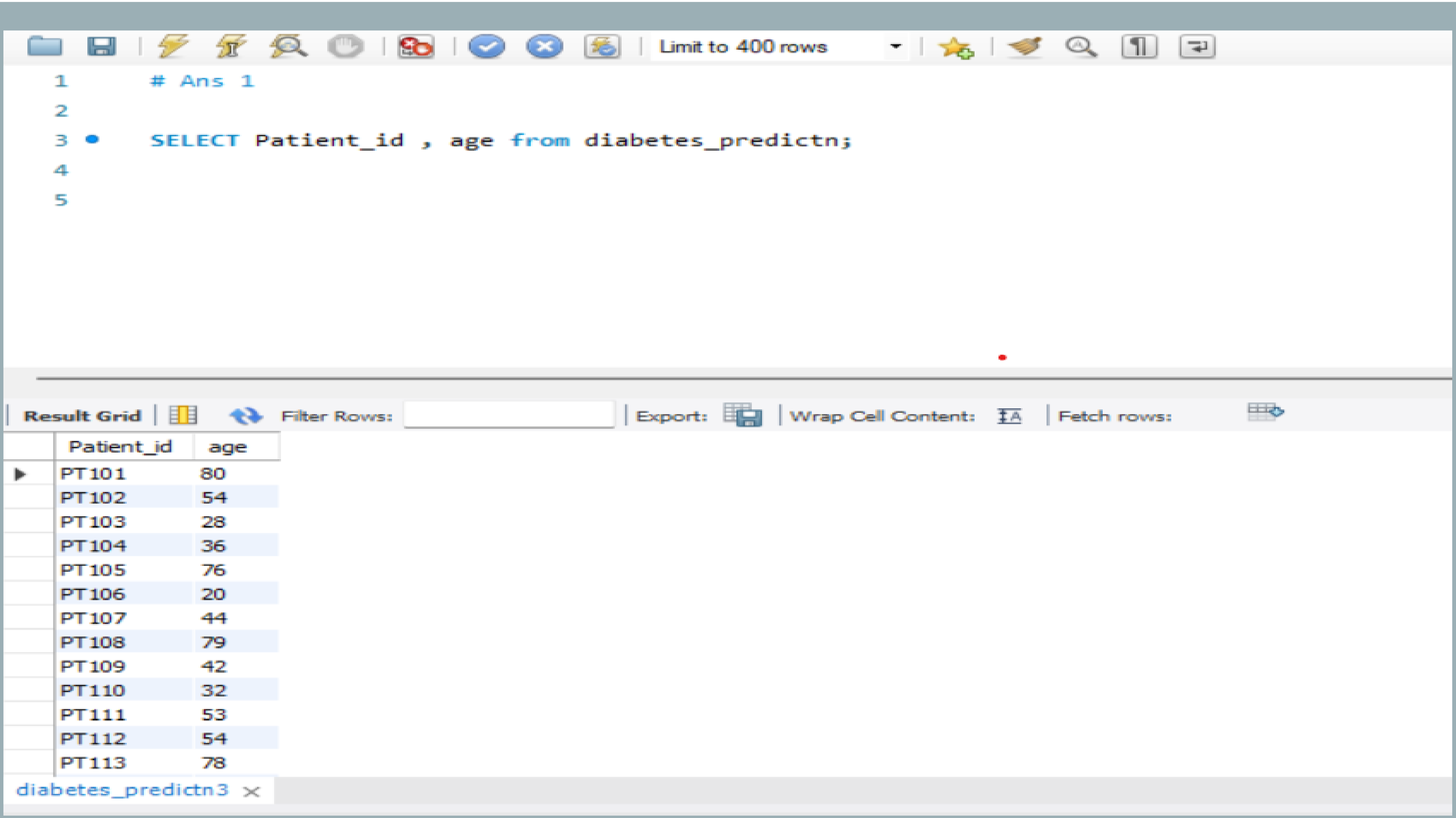
- Now a day Diabetes continues to be a significant global health challenge affecting millions of people worldwide .

- Diabetes is a major cause of blindness, kidney failure, heart attacks, in this dataset we are exploring Diabetes Prediction , Dataset serves as a valuable tool for understanding and predicting the development of diabetes

- The Diabetes Prediction Dataset is a collection of data for predicting the likelihood of an individual developing diabetes based on various attributes such as age, BMI (Body Mass Index), glucose levels, blood pressure, and other health-related features.

- The dataset consist of 1 Lac record and 11 attribute

# 1. Retrieve the Patient_id and ages of all patients.

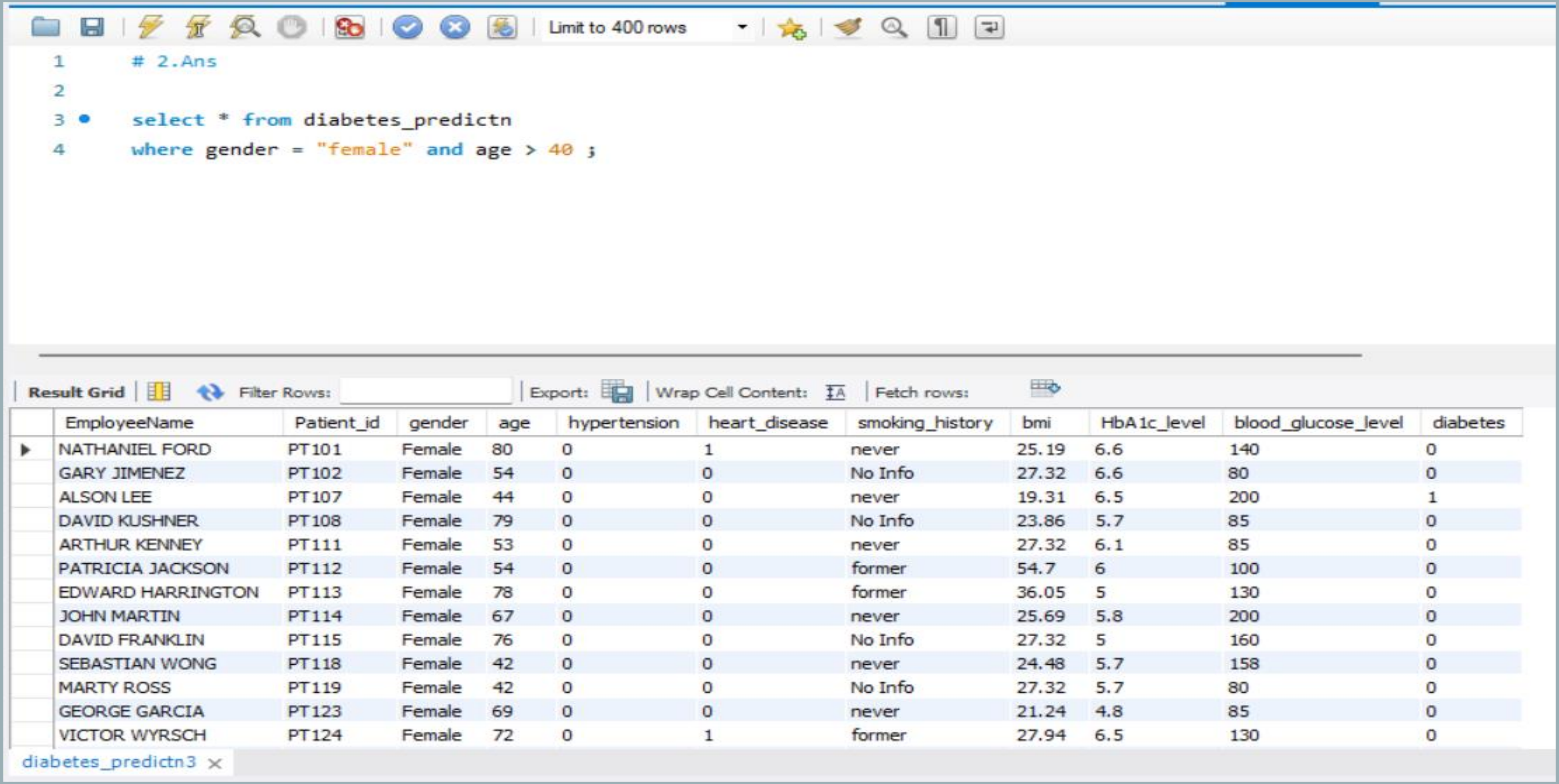


Limit to 400 rows

```
# Ans 1

SELECT Patient_id , age from diabetes_predictn;
```

| Patient_id | age |
|---|---|
| PT101 | 80 |
| PT102 | 54 |
| PT103 | 28 |
| PT104 | 36 |
| PT105 | 76 |
| PT106 | 20 |
| PT107 | 44 |
| PT108 | 79 |
| PT109 | 42 |
| PT110 | 32 |
| PT111 | 53 |
| PT112 | 54 |
| PT113 | 78 |

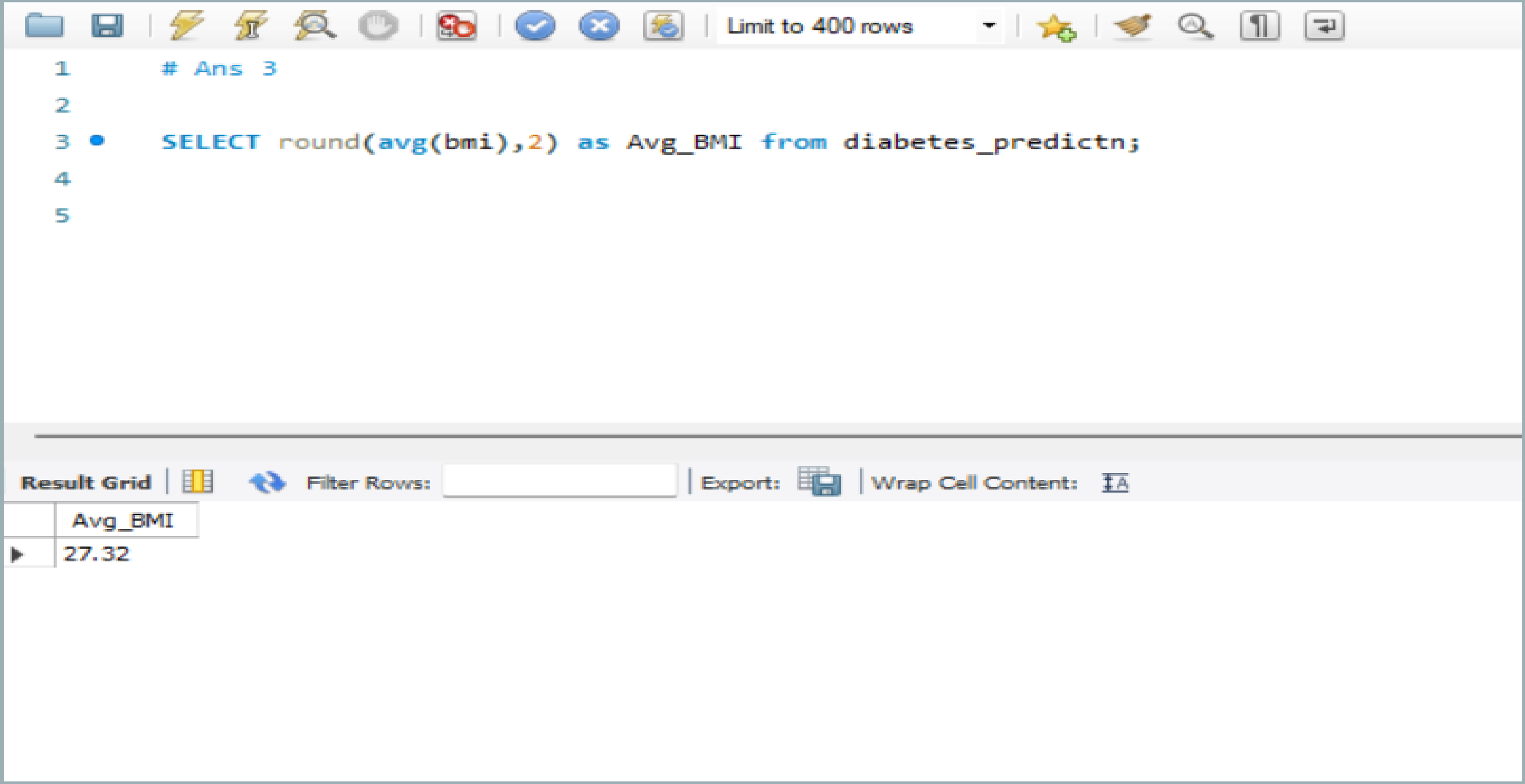
# 2. Select all female patients who are older than 40.



```
# 2.Ans

select * from diabetes_predictn
where gender = "female" and age > 40 ;
```

| 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 |
| DAVID FRANKLIN | PT115 | Female | 76 | 0 | 0 | No Info | 27.32 | 5 | 160 | 0 |
| SEBASTIAN WONG | PT118 | Female | 42 | 0 | 0 | never | 24.48 | 5.7 | 158 | 0 |
| MARTY ROSS | PT119 | Female | 42 | 0 | 0 | No Info | 27.32 | 5.7 | 80 | 0 |
| GEORGE GARCIA | PT123 | Female | 69 | 0 | 0 | never | 21.24 | 4.8 | 85 | 0 |
| VICTOR WYRSCH | PT124 | Female | 72 | 0 | 1 | former | 27.94 | 6.5 | 130 | 0 |

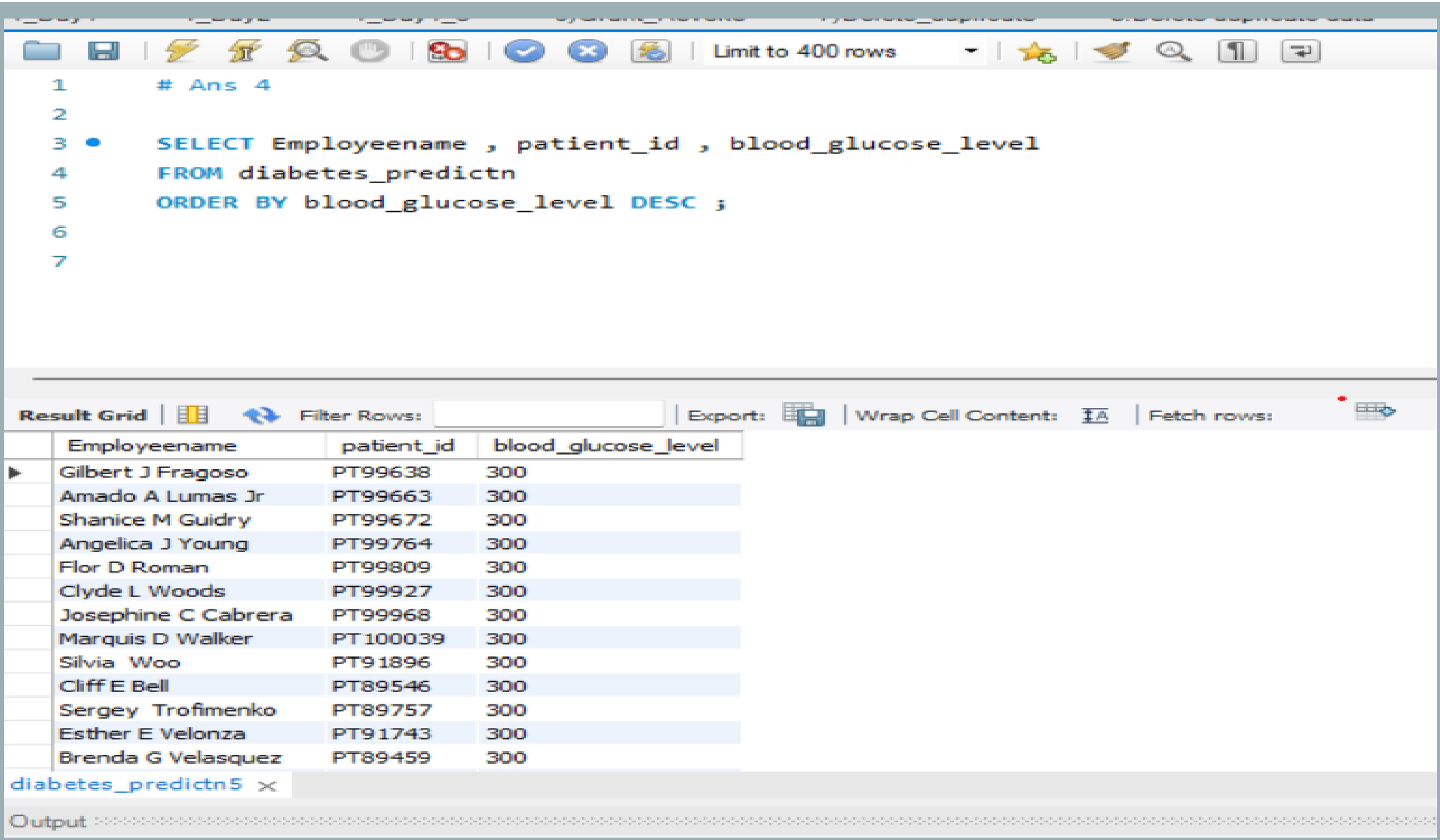
# 3. Calculate the average BMI of patients.



```
# Ans 3

SELECT round(avg(bmi),2) as Avg_BMI from diabetes_predictn;
```

```
Avg_BMI
27.32
```

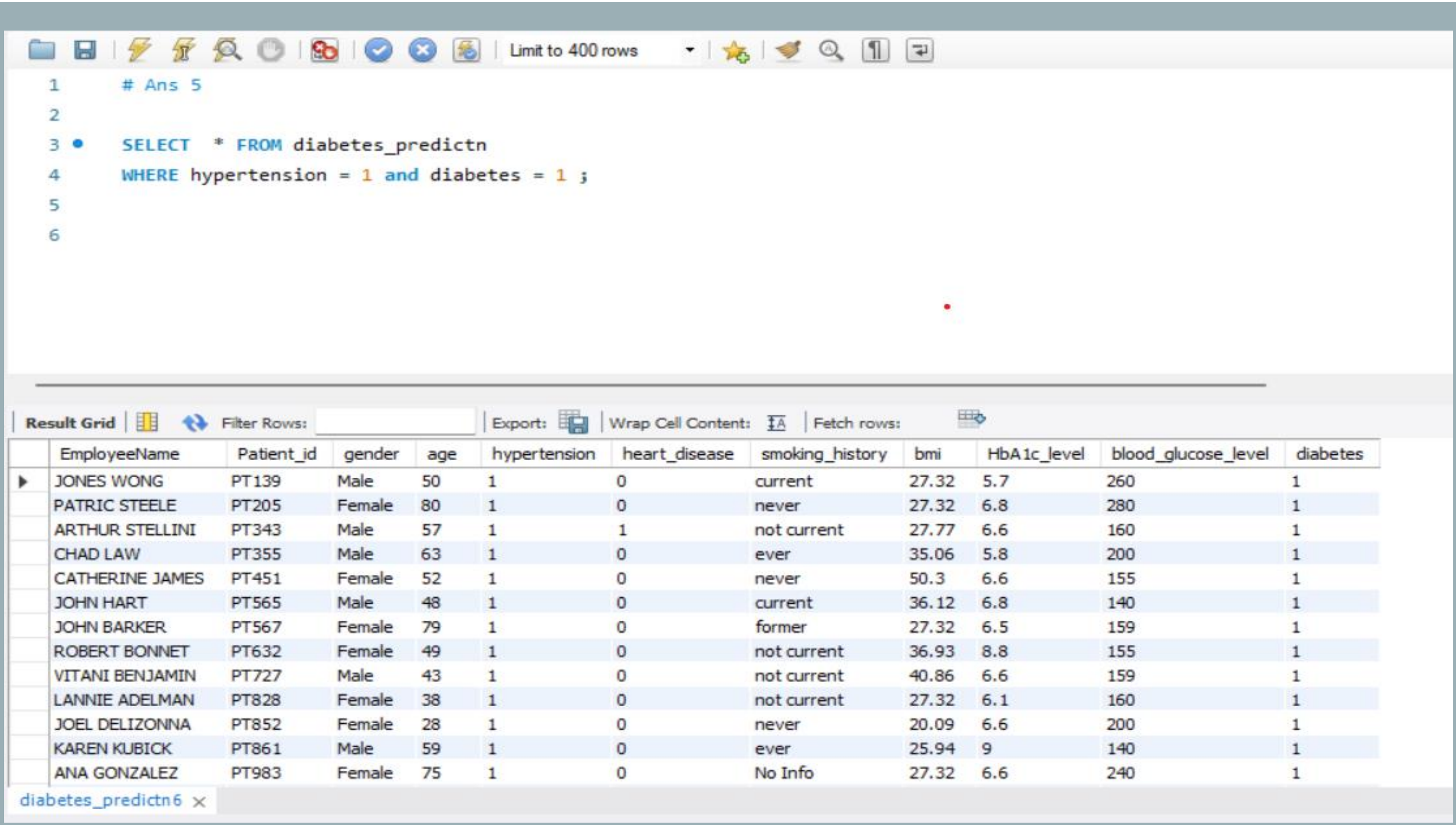
# 4. List patients in descending order of blood glucose levels.



```
# Ans 4

SELECT Employeename , patient_id , blood_glucose_level
FROM diabetes_predictn
ORDER BY blood_glucose_level DESC ;
```

| Employeename | patient_id | blood_glucose_level |
|---|---|---|
| Gilbert J Fragoso | PT99638 | 300 |
| Amado A Lumas Jr | PT99663 | 300 |
| Shanice M Guidry | PT99672 | 300 |
| Angelica J Young | PT99764 | 300 |
| Flor D Roman | PT99809 | 300 |
| Clyde L Woods | PT99927 | 300 |
| Josephine C Cabrera | PT99968 | 300 |
| Marquis D Walker | PT100039 | 300 |
| Silvia Woo | PT91896 | 300 |
| Cliff E Bell | PT89546 | 300 |
| Sergey Trofimenko | PT89757 | 300 |
| Esther E Velonza | PT91743 | 300 |
| Brenda G Velasquez | PT89459 | 300 |

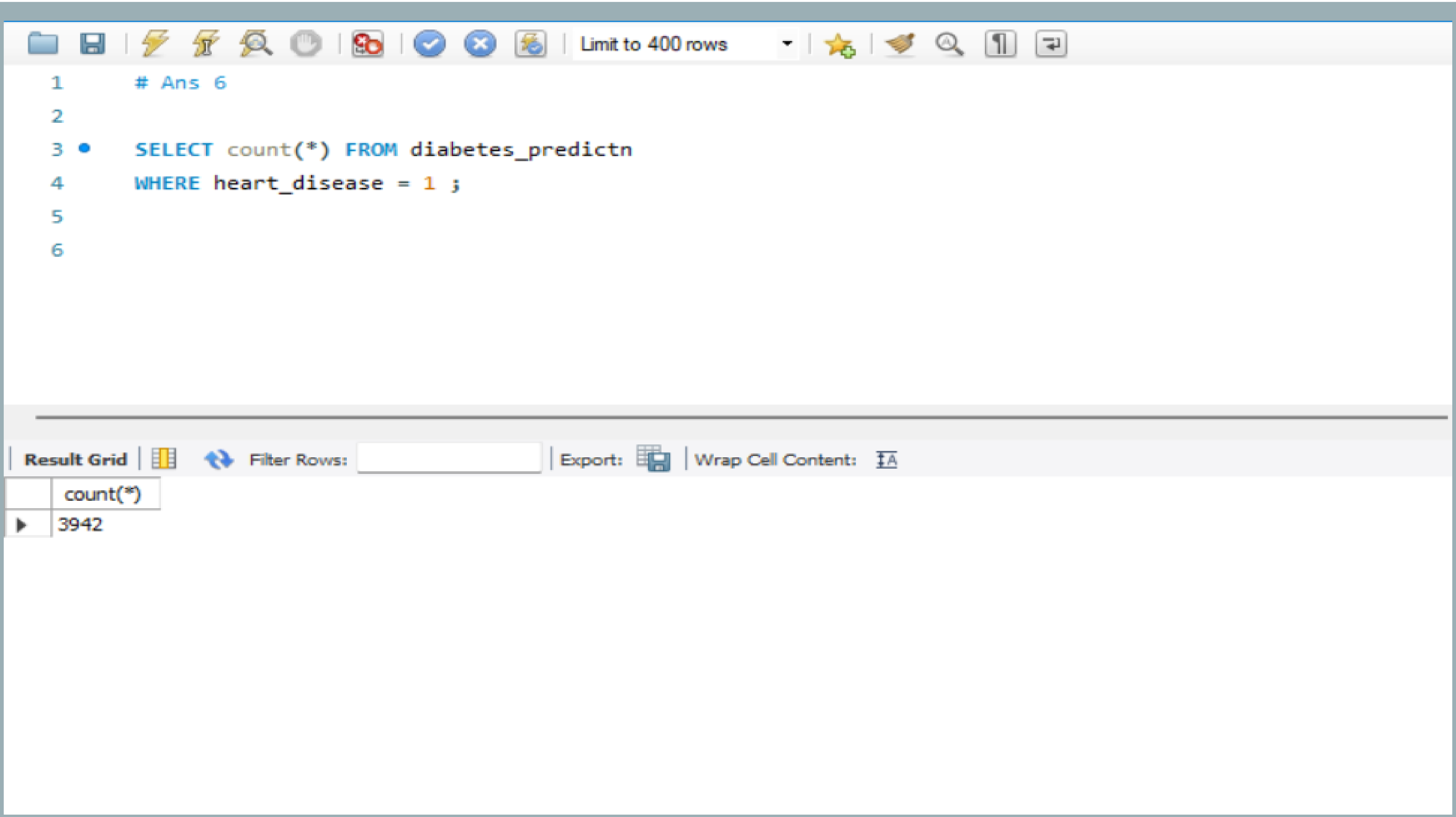
# 5. Find patients who have hypertension and diabetes.



```
# Ans 5

SELECT  * FROM diabetes_predictn
WHERE hypertension = 1 and diabetes = 1 ;
```

| 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 |
| JOHN BARKER | PT567 | Female | 79 | 1 | 0 | former | 27.32 | 6.5 | 159 | 1 |
| ROBERT BONNET | PT632 | Female | 49 | 1 | 0 | not current | 36.93 | 8.8 | 155 | 1 |
| VITANI BENJAMIN | PT727 | Male | 43 | 1 | 0 | not current | 40.86 | 6.6 | 159 | 1 |
| LANNIE ADELMAN | PT828 | Female | 38 | 1 | 0 | not current | 27.32 | 6.1 | 160 | 1 |
| JOEL DELIZONNA | PT852 | Female | 28 | 1 | 0 | never | 20.09 | 6.6 | 200 | 1 |
| KAREN KUBICK | PT861 | Male | 59 | 1 | 0 | ever | 25.94 | 9 | 140 | 1 |
| ANA GONZALEZ | PT983 | Female | 75 | 1 | 0 | No Info | 27.32 | 6.6 | 240 | 1 |

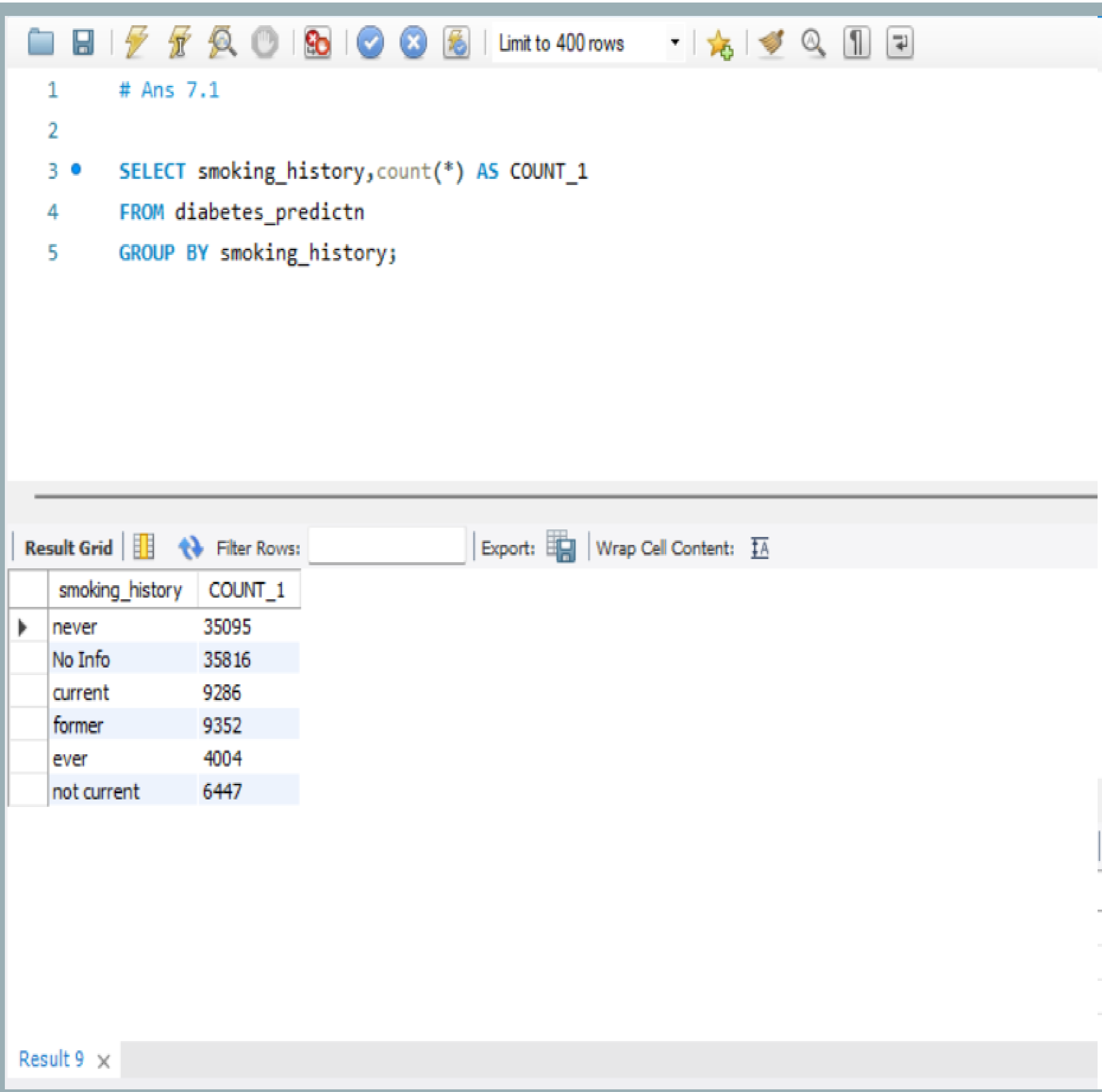
# 6. Determine the number of patients with heart disease.



```
# Ans 6

SELECT count(*) FROM diabetes_predictn
WHERE heart_disease = 1 ;
```

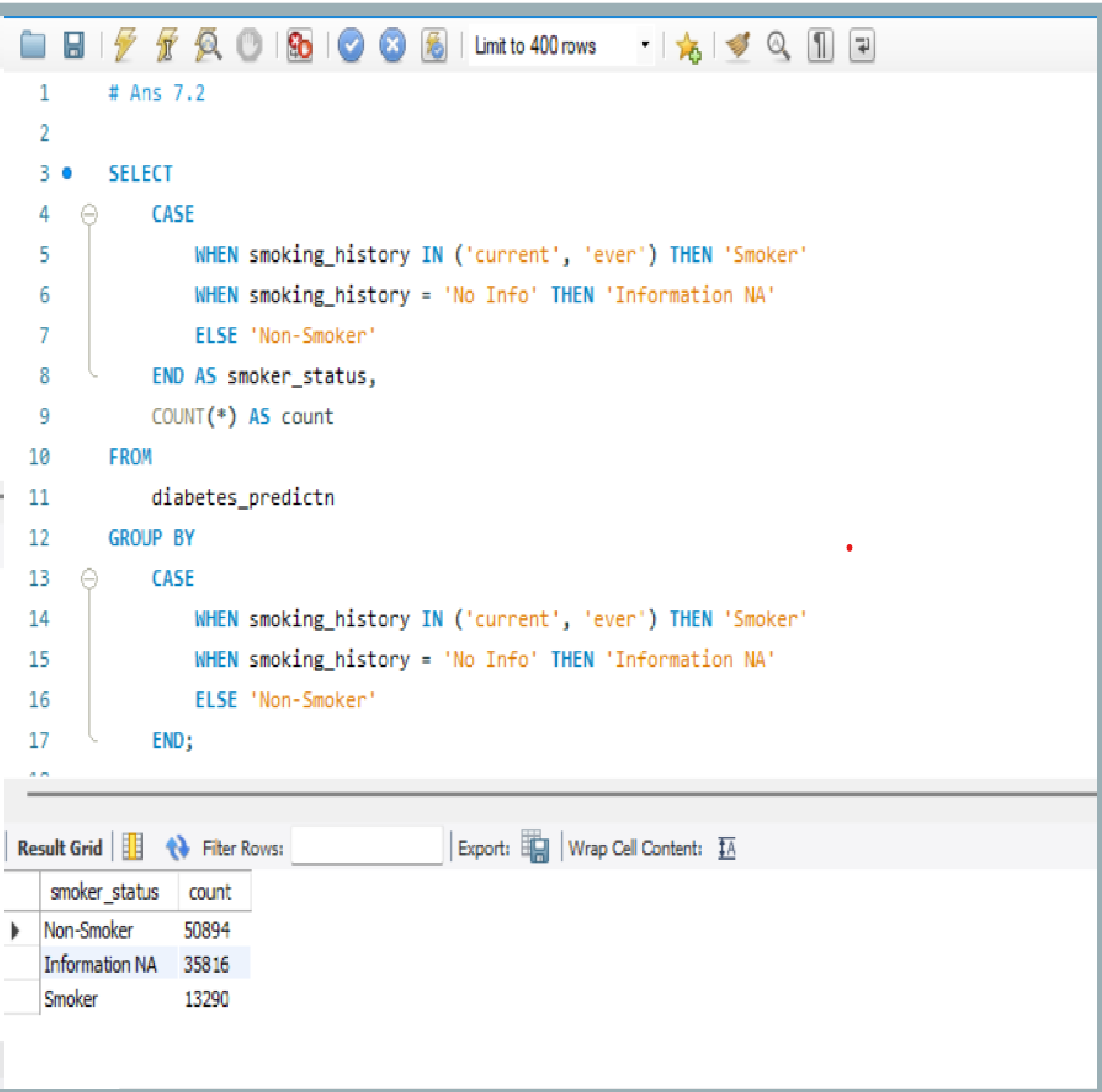
```
count(*)
3942
```

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



```sql
# Ans 7.1

SELECT smoking_history,count(*) AS COUNT_1
FROM diabetes_predictn
GROUP BY smoking_history;
```

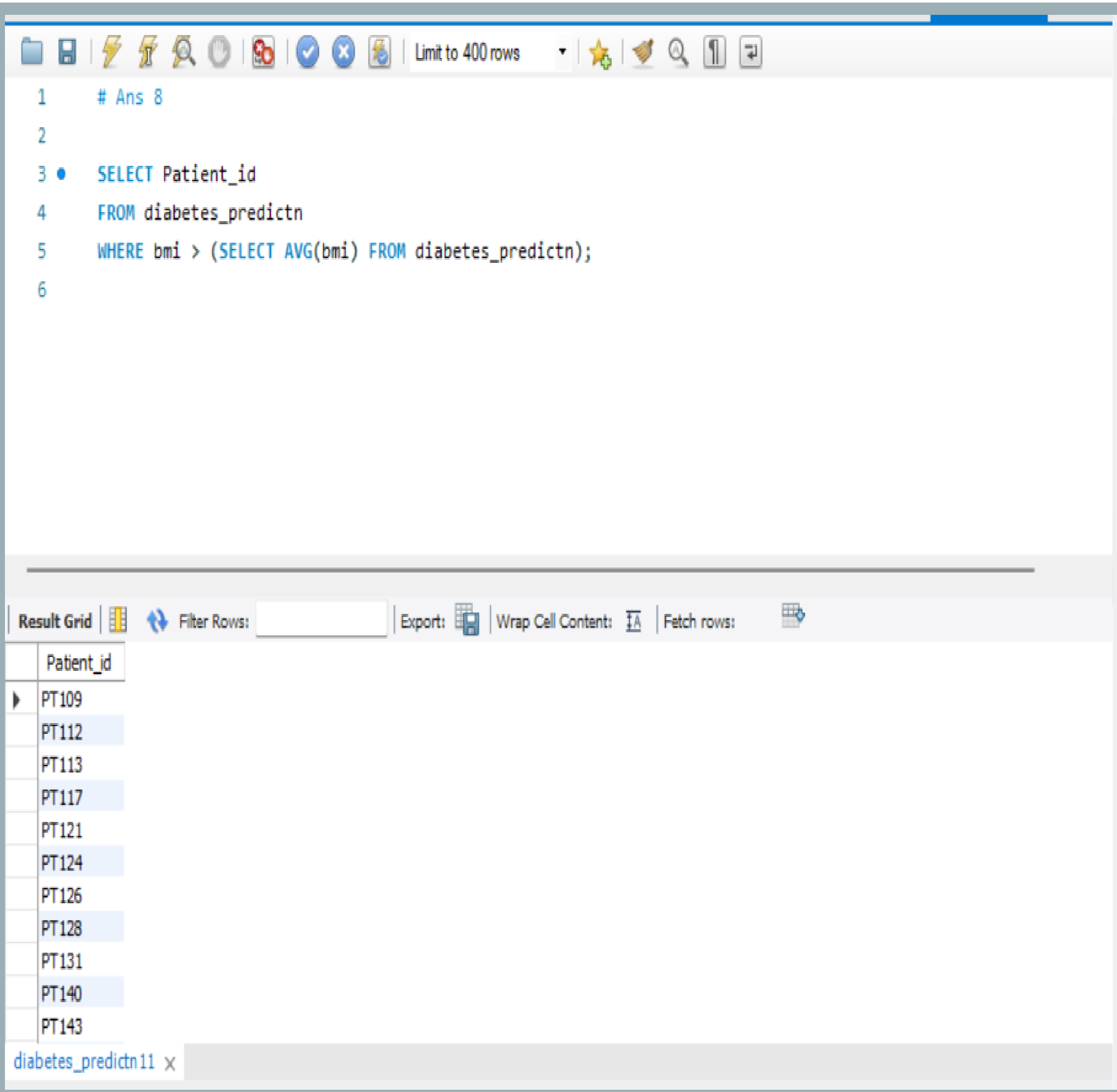
| smoking_history | COUNT_1 |
|---|---|
| never | 35095 |
| No Info | 35816 |
| current | 9286 |
| former | 9352 |
| ever | 4004 |
| not current | 6447 |



```sql
# Ans 7.2

SELECT
    CASE
        WHEN smoking_history IN ('current', 'ever') THEN 'Smoker'
        WHEN smoking_history = 'No Info' THEN 'Information NA'
        ELSE 'Non-Smoker'
    END AS smoker_status,
    COUNT(*) AS count
FROM
    diabetes_predictn
GROUP BY
    CASE
        WHEN smoking_history IN ('current', 'ever') THEN 'Smoker'
        WHEN smoking_history = 'No Info' THEN 'Information NA'
        ELSE 'Non-Smoker'
    END;
```

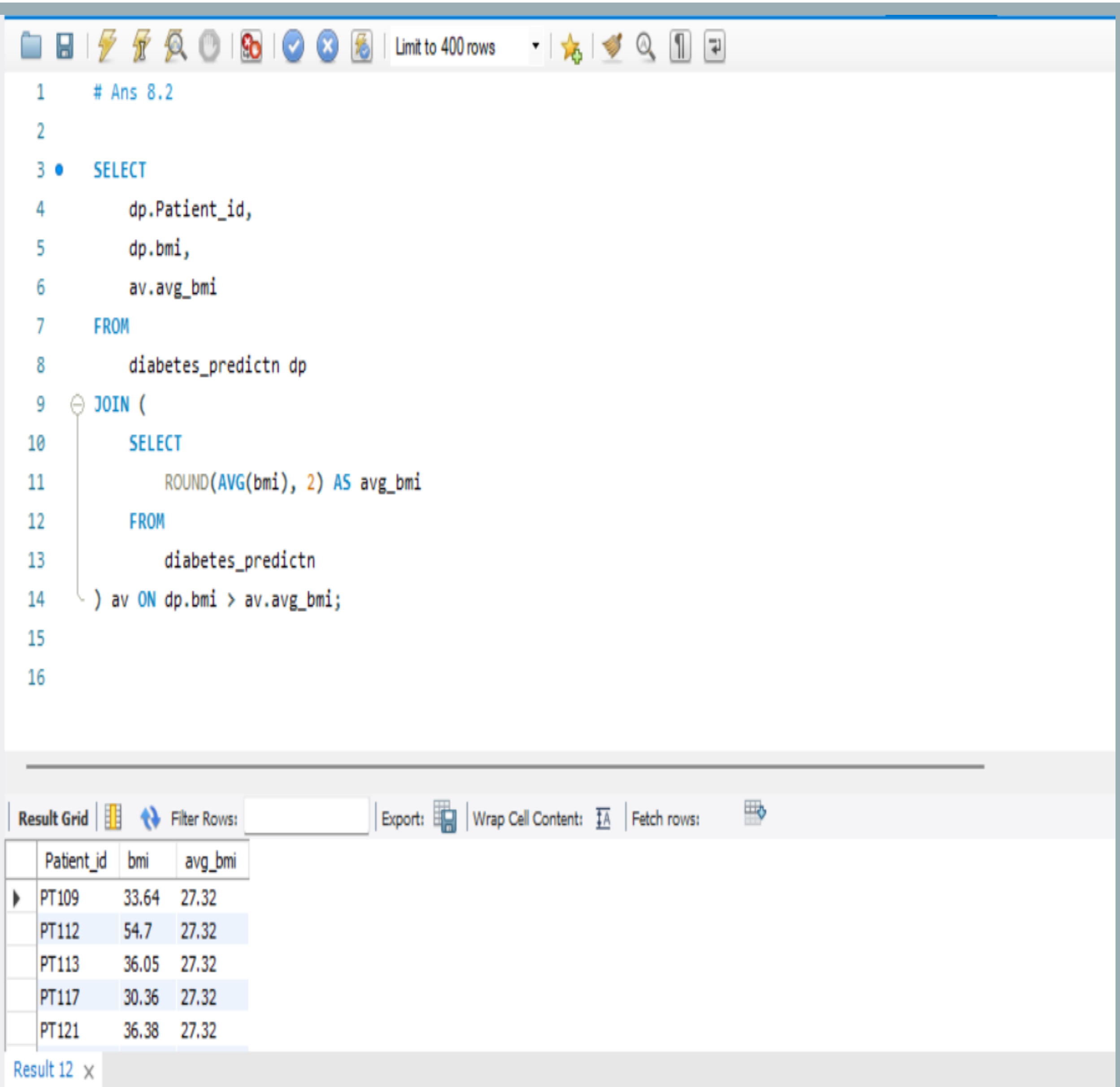
| smoker_status | count |
|---|---|
| Non-Smoker | 50894 |
| Information NA | 35816 |
| Smoker | 13290 |

# 8. Retrieve the Patient_ids of patients who have a BMI greater than the average BMI.



```sql
# Ans 8

SELECT Patient_id
FROM diabetes_predictn
WHERE bmi > (SELECT AVG(bmi) FROM diabetes_predictn);

```

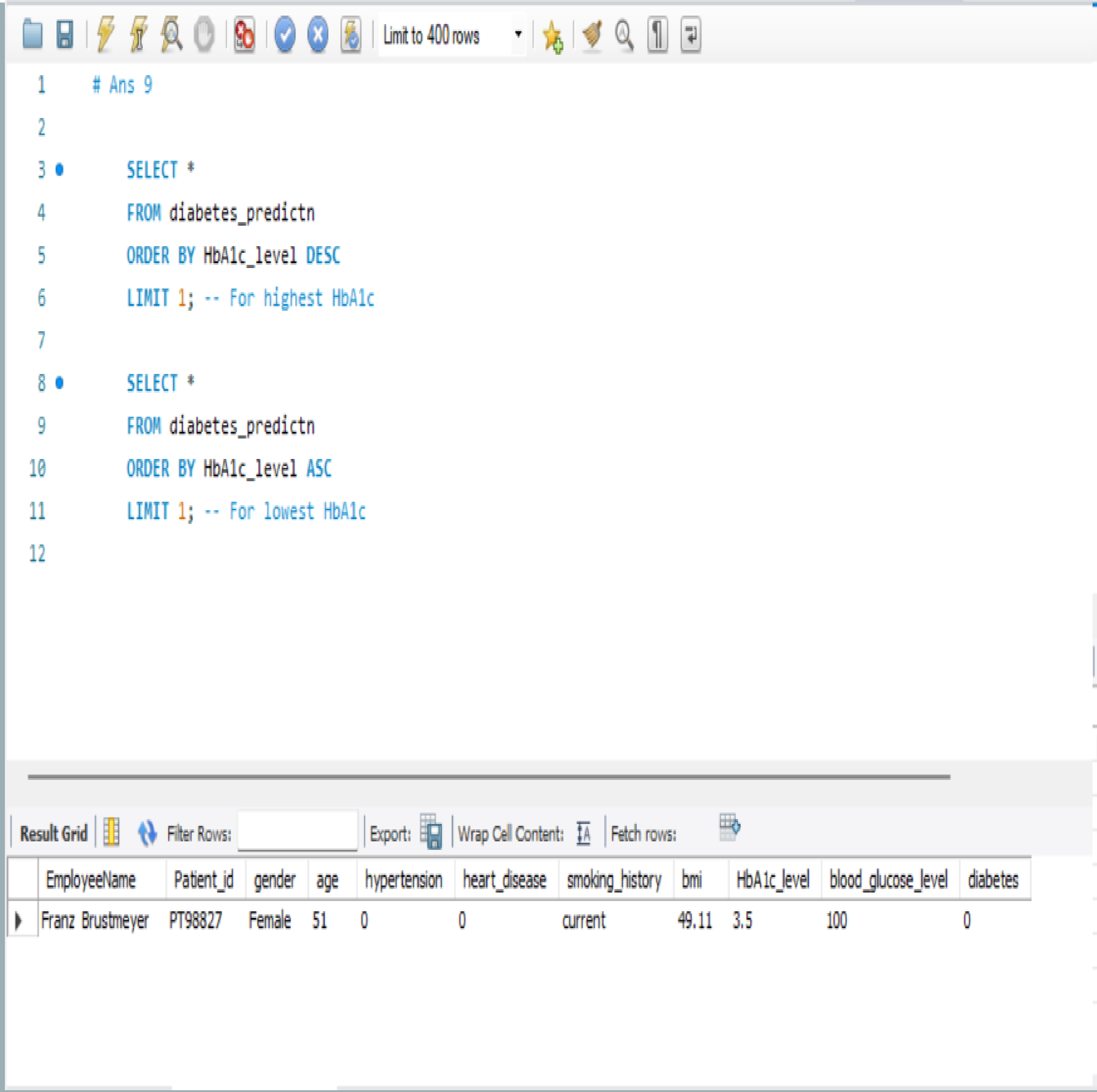
| Patient_id |
|---|
| PT109 |
| PT112 |
| PT113 |
| PT117 |
| PT121 |
| PT124 |
| PT126 |
| PT128 |
| PT131 |
| PT140 |
| PT143 |



```sql
# Ans 8.2

SELECT
    dp.Patient_id,
    dp.bmi,
    av.avg_bmi
FROM
    diabetes_predictn dp
JOIN (
    SELECT
        ROUND(AVG(bmi), 2) AS avg_bmi
    FROM
        diabetes_predictn
) av ON dp.bmi > av.avg_bmi;


```

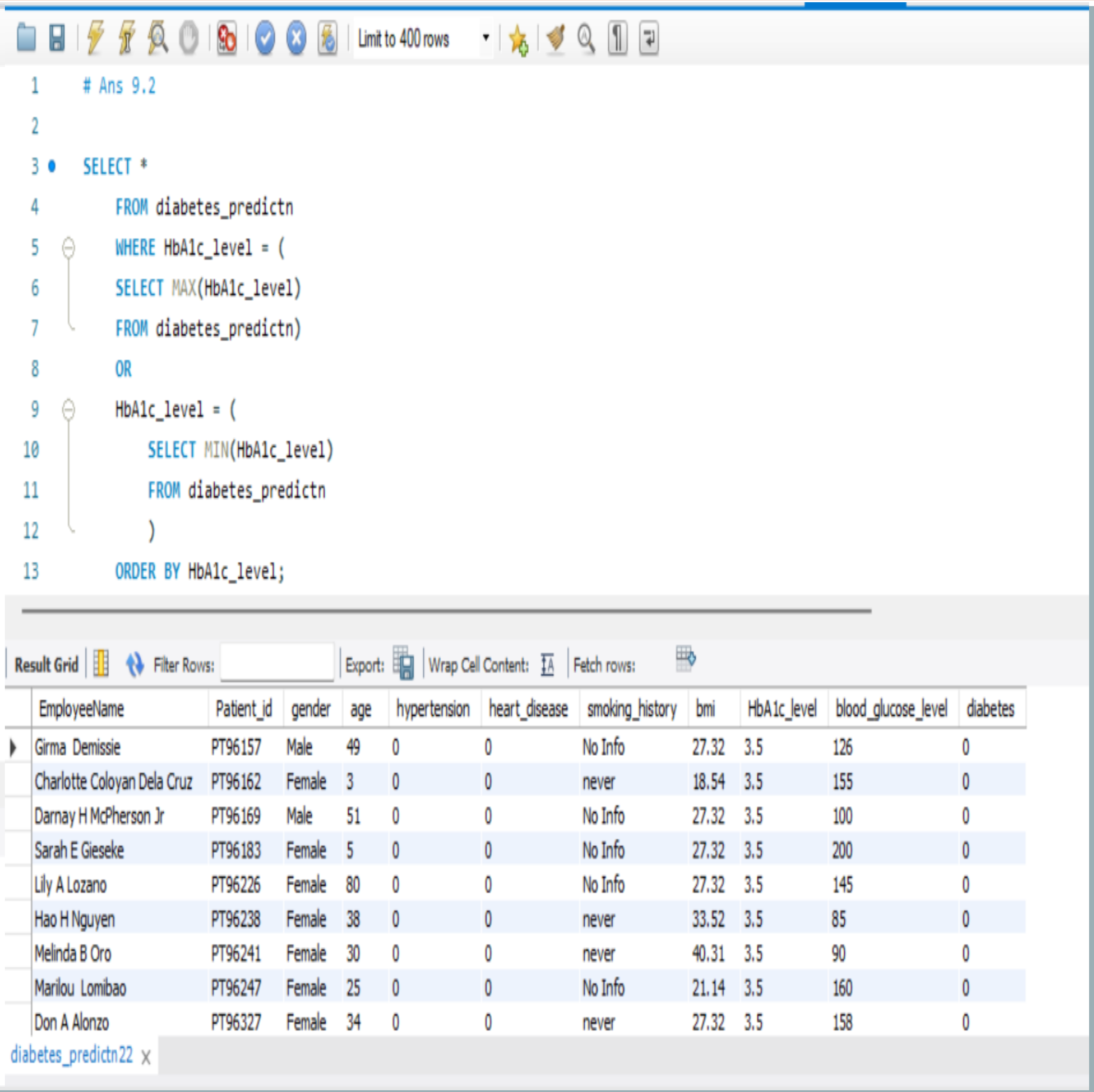
| Patient_id | bmi | avg_bmi |
|---|---|---|
| PT109 | 33.64 | 27.32 |
| PT112 | 54.7 | 27.32 |
| PT113 | 36.05 | 27.32 |
| PT117 | 30.36 | 27.32 |
| PT121 | 36.38 | 27.32 |

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



```
# Ans 9

    SELECT *
    FROM diabetes_predictn
    ORDER BY HbA1c_level DESC
    LIMIT 1; -- For highest HbA1c

    SELECT *
    FROM diabetes_predictn
    ORDER BY HbA1c_level ASC
    LIMIT 1; -- For lowest HbA1c
```

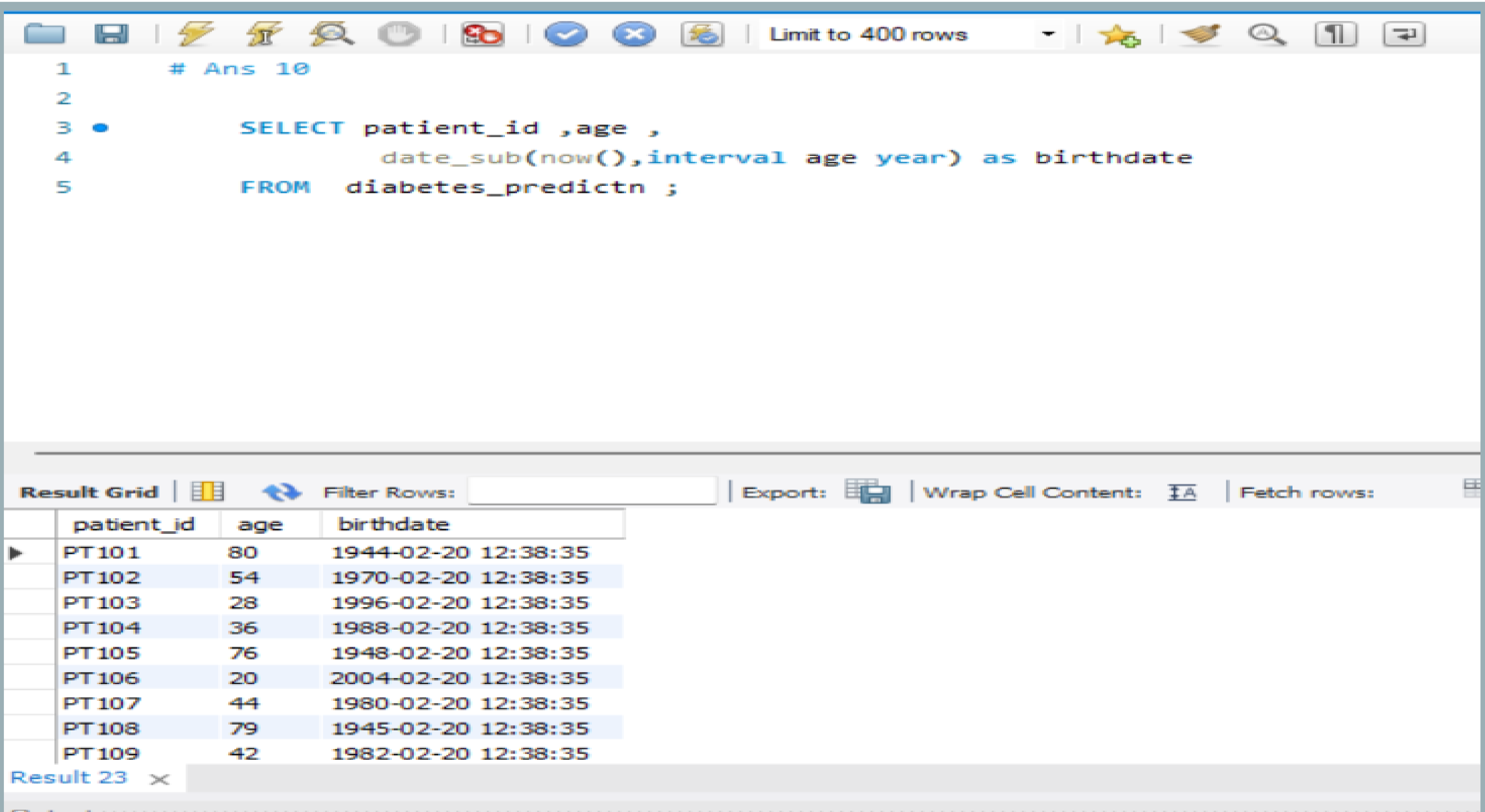
| EmployeeName | Patient_id | gender | age | hypertension | heart_disease | smoking_history | bmi | HbA1c_level | blood_glucose_level | diabetes |
|---|---|---|---|---|---|---|---|---|---|---|
| Franz Brustmeyer | PT98827 | Female | 51 | 0 | 0 | current | 49.11 | 3.5 | 100 | 0 |



```
# Ans 9.2

SELECT *
    FROM diabetes_predictn
    WHERE HbA1c_level = (
    SELECT MAX(HbA1c_level)
    FROM diabetes_predictn)
    OR
    HbA1c_level = (
        SELECT MIN(HbA1c_level)
        FROM diabetes_predictn
        )
    ORDER BY HbA1c_level;
```

| EmployeeName | Patient_id | gender | age | hypertension | heart_disease | smoking_history | bmi | HbA1c_level | blood_glucose_level | diabetes |
|---|---|---|---|---|---|---|---|---|---|---|
| Girma Demissie | PT96157 | Male | 49 | 0 | 0 | No Info | 27.32 | 3.5 | 126 | 0 |
| Charlotte Coloyan Dela Cruz | PT96162 | Female | 3 | 0 | 0 | never | 18.54 | 3.5 | 155 | 0 |
| Darnay H McPherson Jr | PT96169 | Male | 51 | 0 | 0 | No Info | 27.32 | 3.5 | 100 | 0 |
| Sarah E Gieseke | PT96183 | Female | 5 | 0 | 0 | No Info | 27.32 | 3.5 | 200 | 0 |
| Lily A Lozano | PT96226 | Female | 80 | 0 | 0 | No Info | 27.32 | 3.5 | 145 | 0 |
| Hao H Nguyen | PT96238 | Female | 38 | 0 | 0 | never | 33.52 | 3.5 | 85 | 0 |
| Melinda B Oro | PT96241 | Female | 30 | 0 | 0 | never | 40.31 | 3.5 | 90 | 0 |
| Marilou Lomibao | PT96247 | Female | 25 | 0 | 0 | No Info | 21.14 | 3.5 | 160 | 0 |
| Don A Alonzo | PT96327 | Female | 34 | 0 | 0 | never | 27.32 | 3.5 | 158 | 0 |

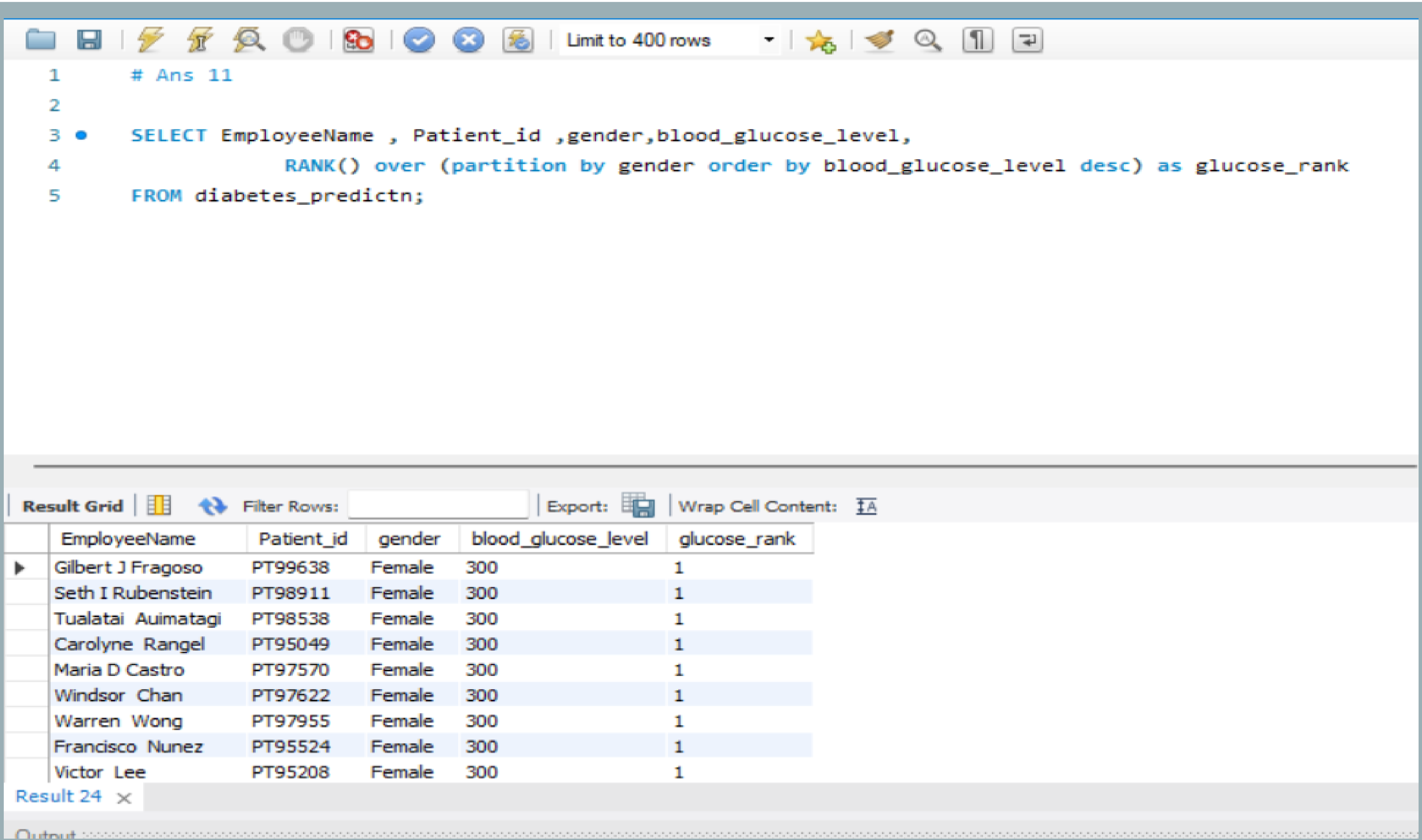
# 10. Calculate the age of patients in years (assuming the current date as of now).



```sql
# Ans 10

        SELECT patient_id ,age ,
                date_sub(now(),interval age year) as birthdate
        FROM  diabetes_predictn ;
```

| patient_id | age | birthdate |
|---|---|---|
| PT101 | 80 | 1944-02-20 12:38:35 |
| PT102 | 54 | 1970-02-20 12:38:35 |
| PT103 | 28 | 1996-02-20 12:38:35 |
| PT104 | 36 | 1988-02-20 12:38:35 |
| PT105 | 76 | 1948-02-20 12:38:35 |
| PT106 | 20 | 2004-02-20 12:38:35 |
| PT107 | 44 | 1980-02-20 12:38:35 |
| PT108 | 79 | 1945-02-20 12:38:35 |
| PT109 | 42 | 1982-02-20 12:38:35 |

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

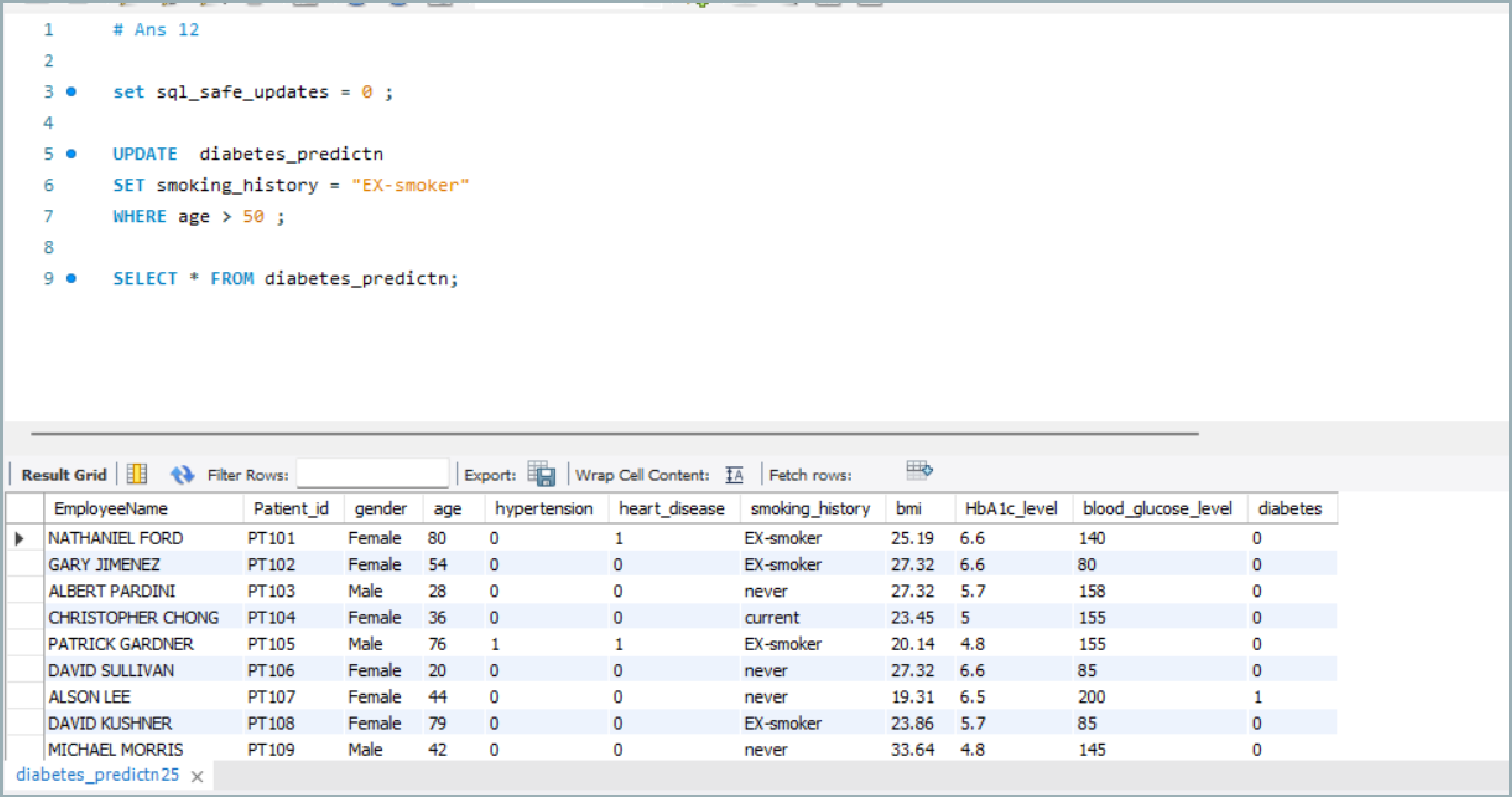


Limit to 400 rows

```sql
# Ans 11

SELECT EmployeeName , Patient_id ,gender,blood_glucose_level,
          RANK() over (partition by gender order by blood_glucose_level desc) as glucose_rank
FROM diabetes_predictn;
```

| EmployeeName | Patient_id | gender | blood_glucose_level | glucose_rank |
|---|---|---|---|---|
| Gilbert J Fragoso | PT99638 | Female | 300 | 1 |
| Seth I Rubenstein | PT98911 | Female | 300 | 1 |
| Tualatai  Auimatagi | PT98538 | Female | 300 | 1 |
| Carolyne  Rangel | PT95049 | Female | 300 | 1 |
| Maria D Castro | PT97570 | Female | 300 | 1 |
| Windsor  Chan | PT97622 | Female | 300 | 1 |
| Warren  Wong | PT97955 | Female | 300 | 1 |
| Francisco  Nunez | PT95524 | Female | 300 | 1 |
| Victor  Lee | PT95208 | Female | 300 | 1 |

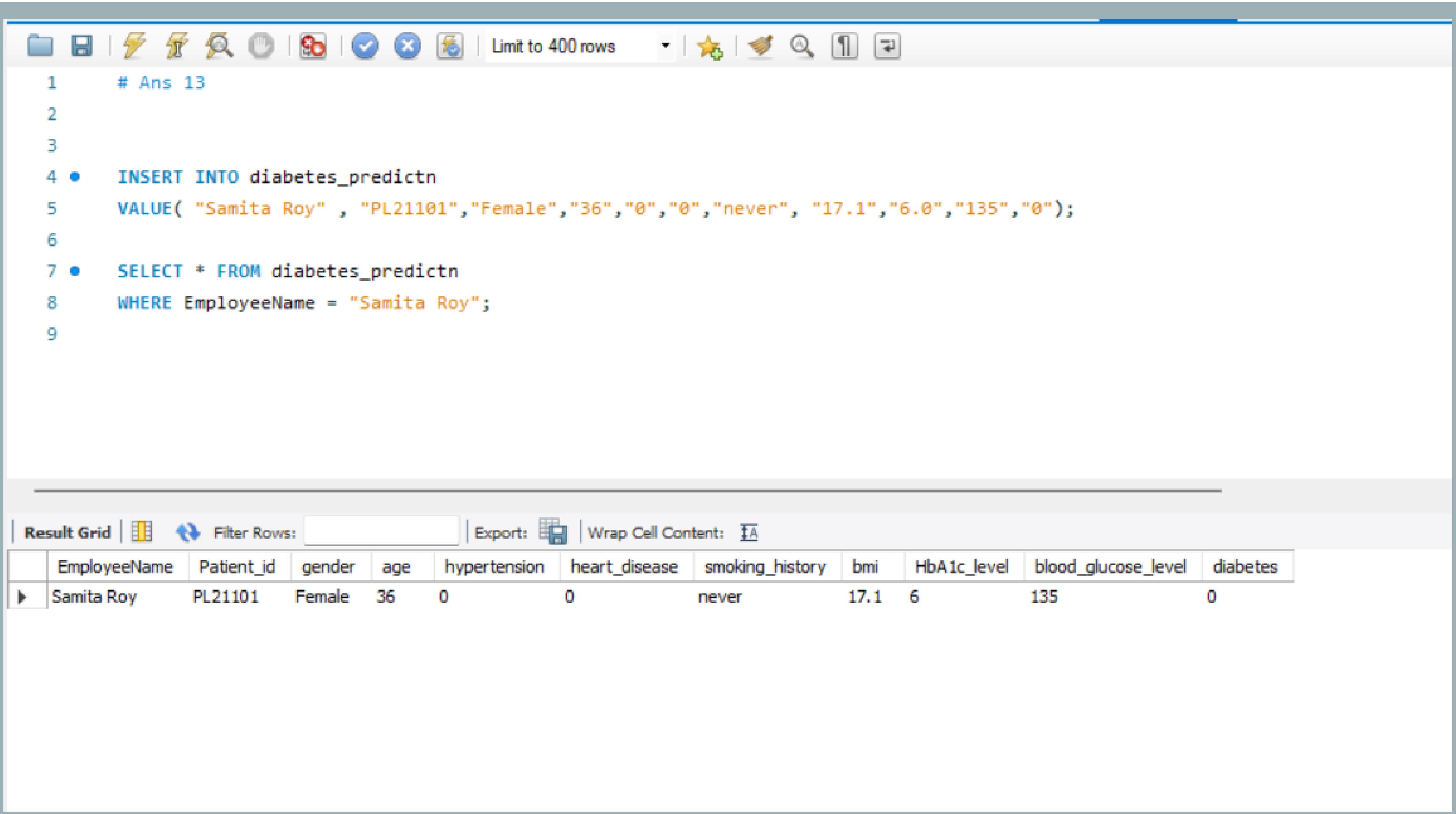
# 12. Update the smoking history of patients who are older than 50 to "Ex-smoker."



```
# Ans 12

set sql_safe_updates = 0 ;

UPDATE  diabetes_predictn
SET smoking_history = "EX-smoker"
WHERE age > 50 ;

SELECT * FROM diabetes_predictn;
```

| 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 |
| 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 |

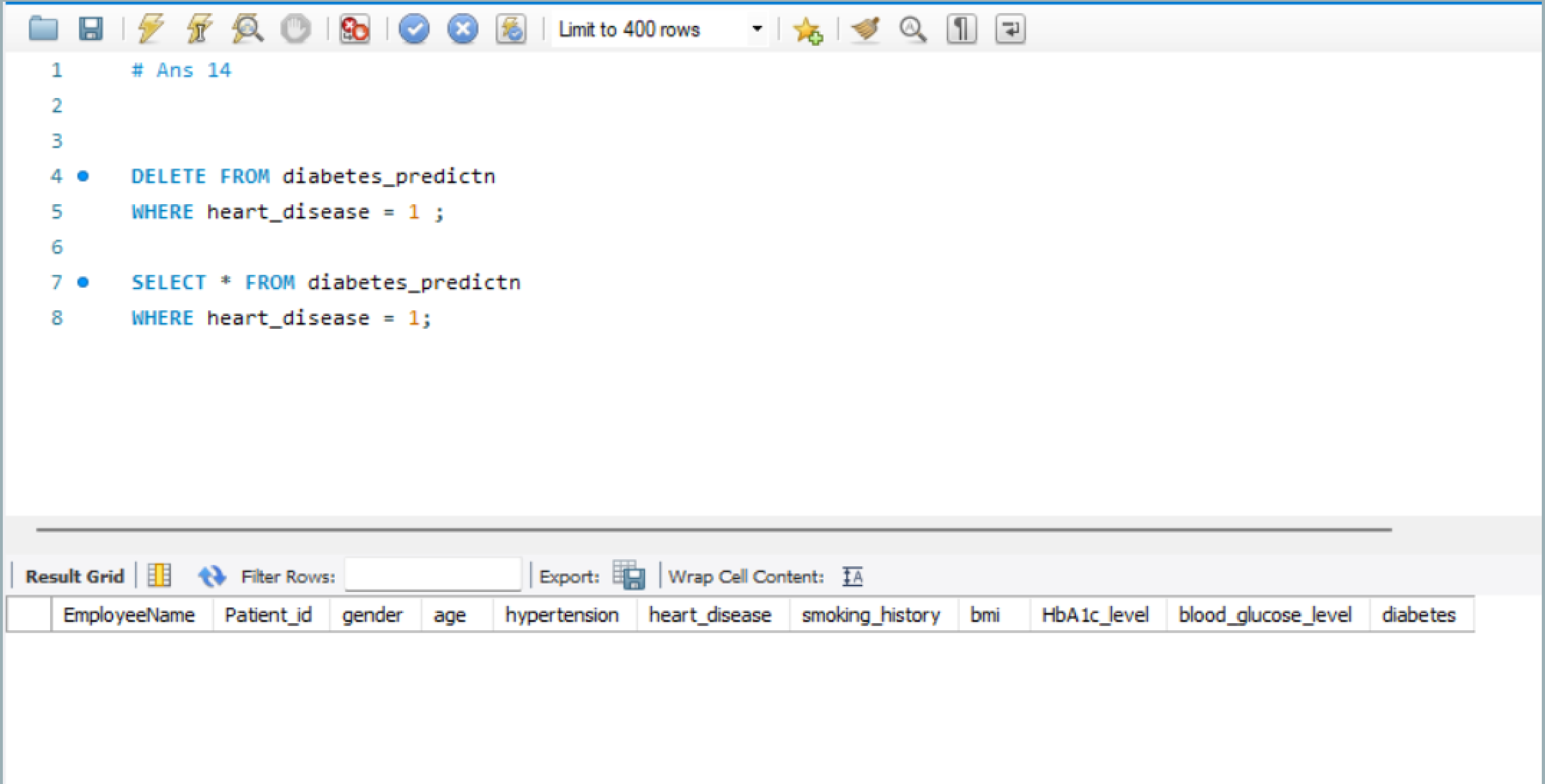
# 13. Insert a new patient into the database with sample data.



```
# Ans 13


INSERT INTO diabetes_predictn
VALUE( "Samita Roy" , "PL21101","Female","36","0","0","never", "17.1","6.0","135","0");

SELECT * FROM diabetes_predictn
WHERE EmployeeName = "Samita Roy";

```

| EmployeeName | Patient_id | gender | age | hypertension | heart_disease | smoking_history | bmi | HbA1c_level | blood_glucose_level | diabetes |
|---|---|---|---|---|---|---|---|---|---|---|
| Samita Roy | PL21101 | Female | 36 | 0 | 0 | never | 17.1 | 6 | 135 | 0 |

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

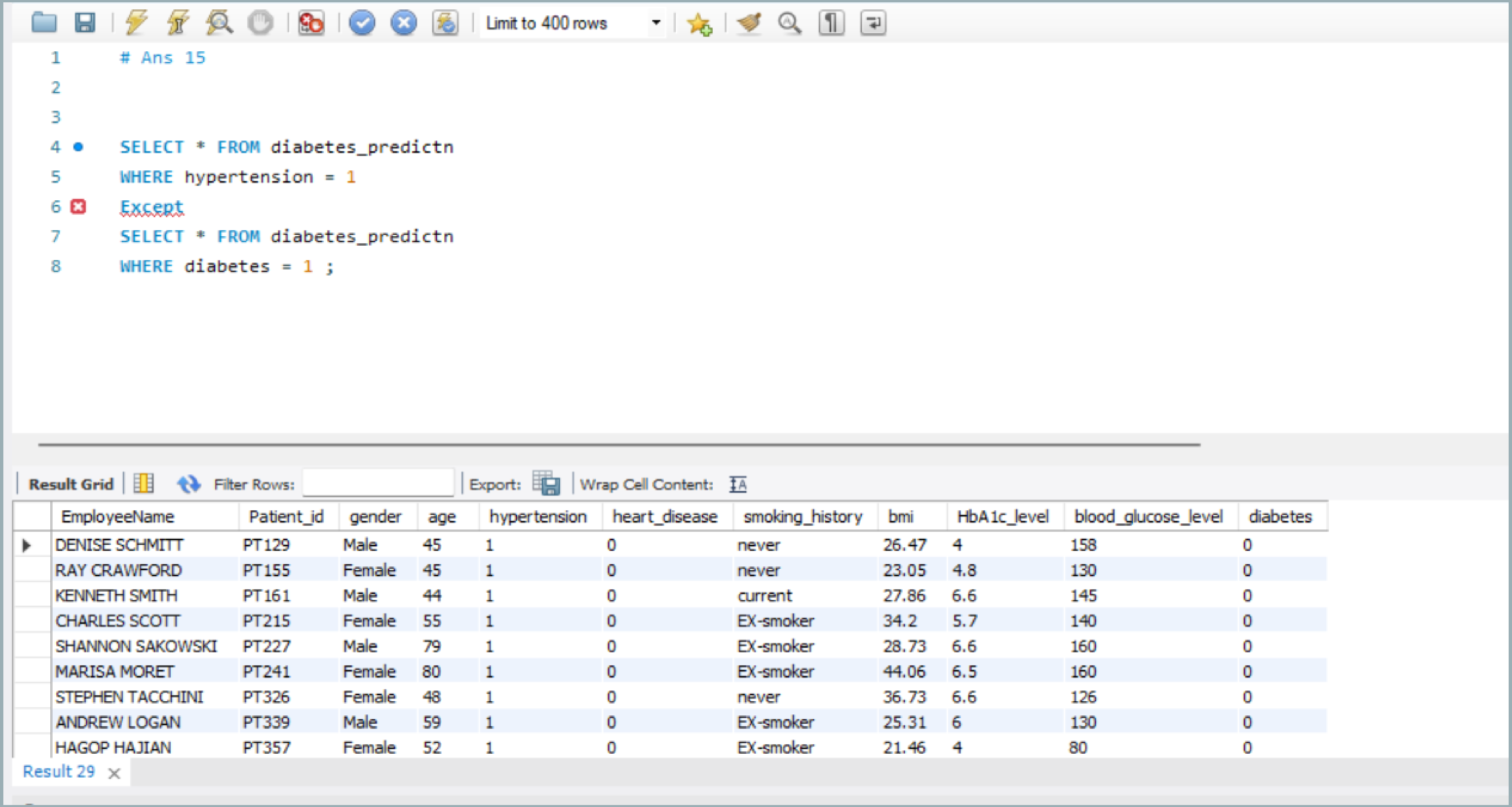


Limit to 400 rows

```
# Ans 14


DELETE FROM diabetes_predictn
WHERE heart_disease = 1 ;

SELECT * FROM diabetes_predictn
WHERE heart_disease = 1;
```

| | EmployeeName | Patient_id | gender | age | hypertension | heart_disease | smoking_history | bmi | HbA1c_level | blood_glucose_level | diabetes |
|---|---|---|---|---|---|---|---|---|---|---|---|

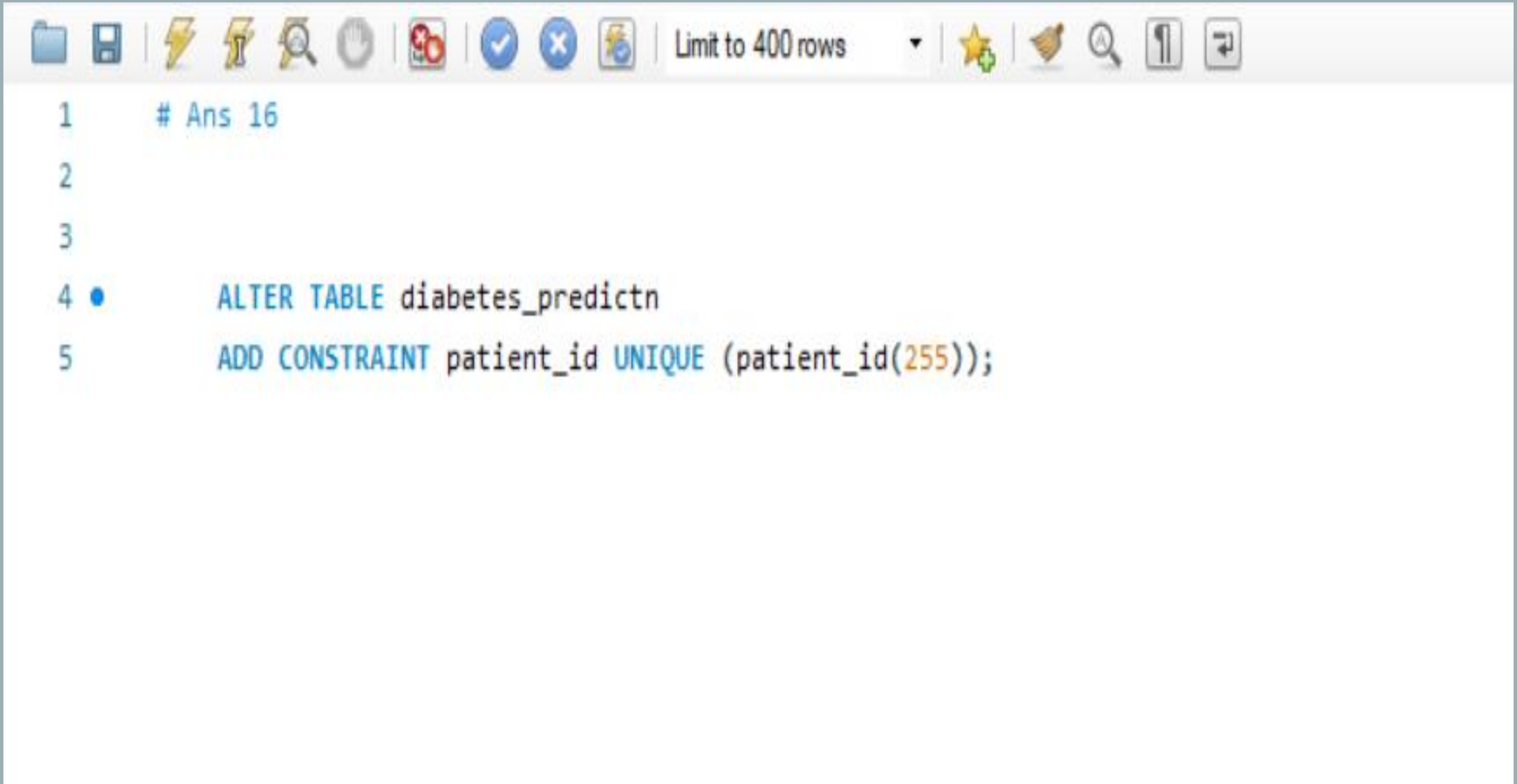
# 15. Find patients who have hypertension but not diabetes using the EXCEPT operator.



```
# Ans 15


SELECT * FROM diabetes_predictn
WHERE hypertension = 1
Except
SELECT * FROM diabetes_predictn
WHERE diabetes = 1 ;
```

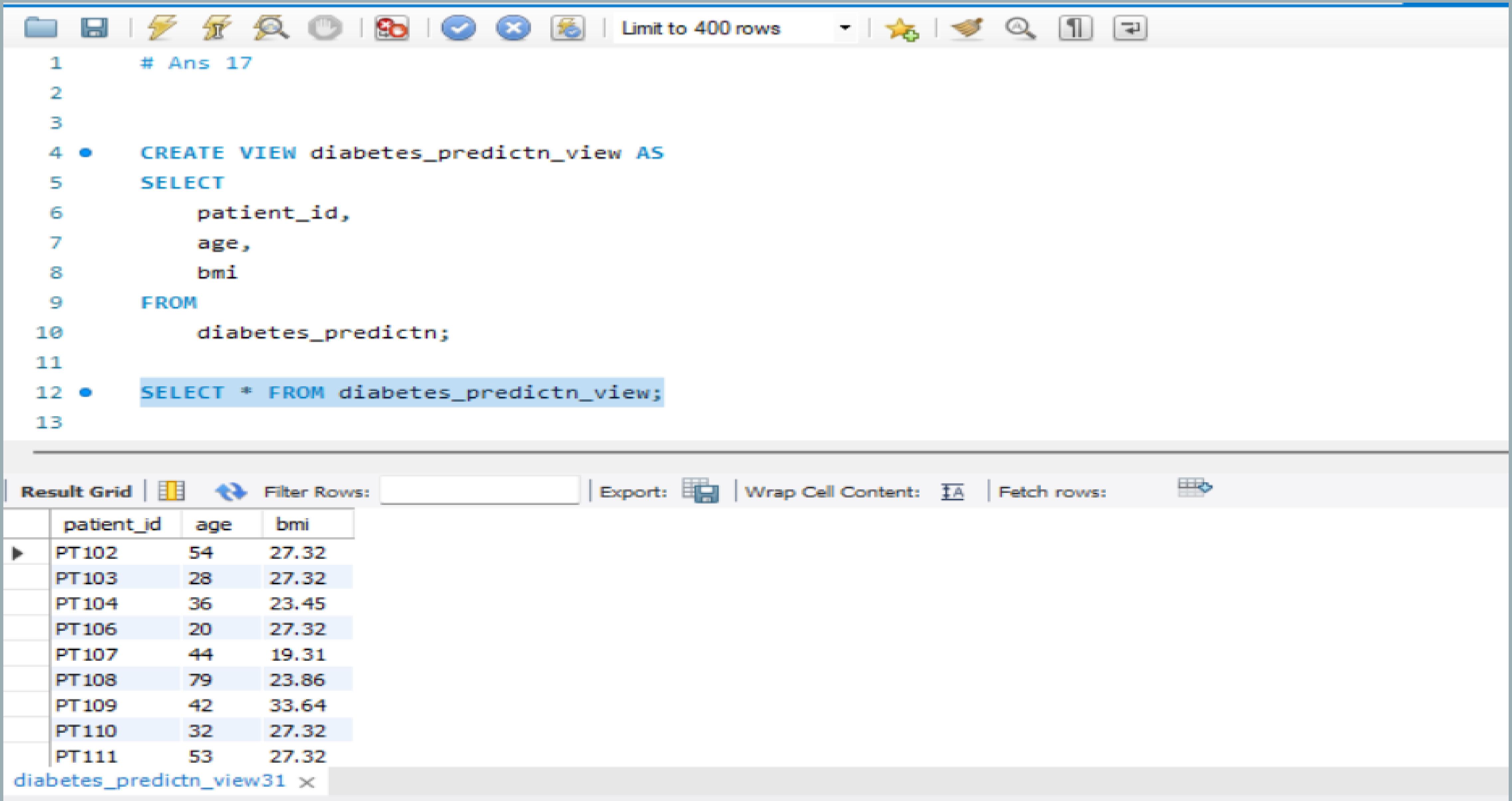
| EmployeeName | Patient_id | gender | age | hypertension | heart_disease | smoking_history | bmi | HbA1c_level | blood_glucose_level | diabetes |
|---|---|---|---|---|---|---|---|---|---|---|
| DENISE SCHMITT | PT129 | Male | 45 | 1 | 0 | never | 26.47 | 4 | 158 | 0 |
| RAY CRAWFORD | PT155 | Female | 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 |
| ANDREW LOGAN | PT339 | Male | 59 | 1 | 0 | EX-smoker | 25.31 | 6 | 130 | 0 |
| HAGOP HAJIAN | PT357 | Female | 52 | 1 | 0 | EX-smoker | 21.46 | 4 | 80 | 0 |

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



```
# Ans 16


    ALTER TABLE diabetes_predictn
    ADD CONSTRAINT patient_id UNIQUE (patient_id(255));
```

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



```
# Ans 17


CREATE VIEW diabetes_predictn_view AS
SELECT
    patient_id,
    age,
    bmi
FROM
    diabetes_predictn;

SELECT * FROM diabetes_predictn_view;
```

| patient_id | age | bmi |
|---|---|---|
| PT102 | 54 | 27.32 |
| PT103 | 28 | 27.32 |
| PT104 | 36 | 23.45 |
| PT106 | 20 | 27.32 |
| PT107 | 44 | 19.31 |
| PT108 | 79 | 23.86 |
| PT109 | 42 | 33.64 |
| PT110 | 32 | 27.32 |
| PT111 | 53 | 27.32 |

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

1) **Normalization: Break down large tables into smaller**

2) **Use of Primary Keys: It helps in maintaining data integrity and avoiding duplicate records.**

3) **Foreign Key Constraints: Implement foreign key constraints to enforce referential integrity between related tables**

4) **Unique Constraints: Apply unique constraints to prevent duplicate values within columns where necessary, ensuring data consistency.**

5) **Regular Maintenance: Regularly review and update the database schema based on changing requirements and performance considerations.**

6) **Review Data Types: Use appropriate data types for columns to optimize storage and ensure data integrity. Avoid using excessively large data types where not necessary.**

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

1) **Write efficient queries with proper filtering and minimal calculations.**

2) **Duplicate data selectively to reduce complexity and speed up queries**

3) **Avoid SELECT * Instead of selecting all columns using SELECT *, explicitly list the required columns.This reduces the amount of data transferred between the database server and the client application, resulting in faster query execution.**

4) **Retrieve only the necessary columns and rows from the database by using SELECT**

5) **Reduce the use of subqueries .Prefer EXISTS and IN operators over subqueries or JOINs when checking for the existence of records in another table**

Thanking you