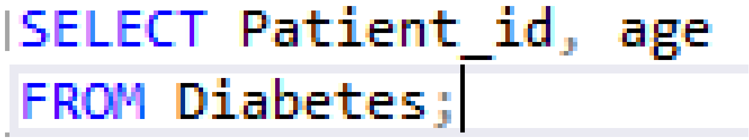
DIABETES PREDICTION

PSYLIQ DATA ANALYST INTERNSHIP

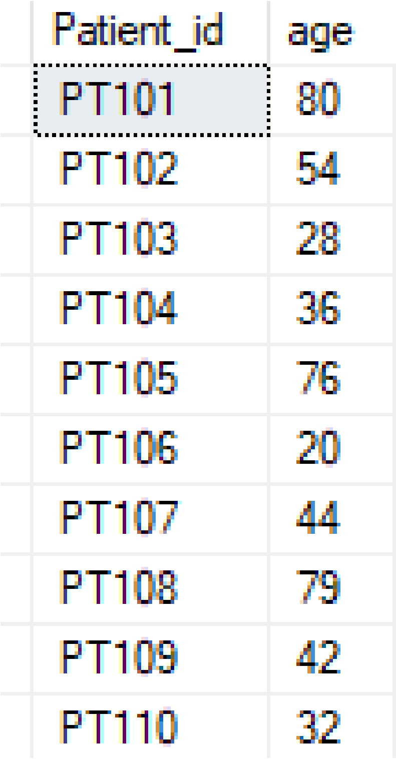
LILIANA MORONES ALBA

TASK 2

1. Retrieve the Patient_id and ages of all patients.

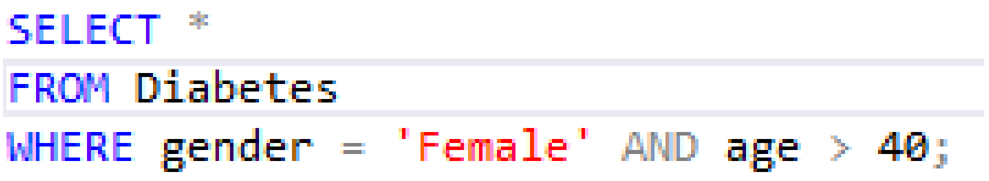


```sql
SELECT Patient_id, age
FROM Diabetes;
```

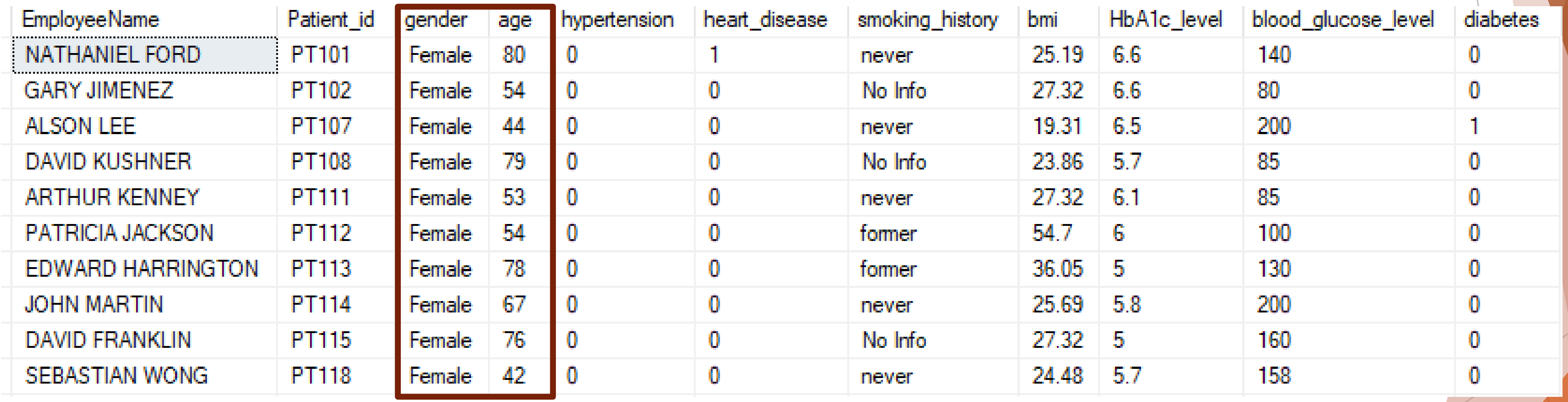


Patient_id	age
PT101	80
PT102	54
PT103	28
PT104	36
PT105	76
PT106	20
PT107	44
PT108	79
PT109	42
PT110	32

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

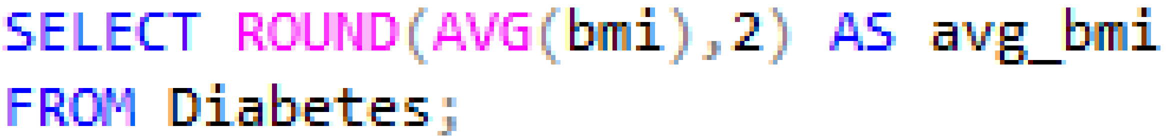


```
SELECT *
FROM Diabetes
WHERE gender = 'Female' AND age > 40;
```

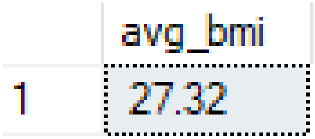


EmployeeName	Patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
NATHANIEL FORD	PT101	Female	80	0	1	never	25.19	6.6	140	0
GARY JIMENEZ	PT102	Female	54	0	0	No Info	27.32	6.6	80	0
ALSON LEE	PT107	Female	44	0	0	never	19.31	6.5	200	1
DAVID KUSHNER	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

3. Calculate the average BMI of patients.

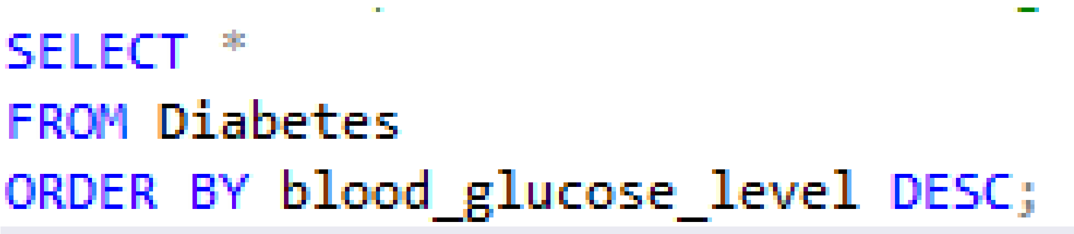


```sql
SELECT ROUND(AVG(bmi),2) AS avg_bmi
FROM Diabetes;
```

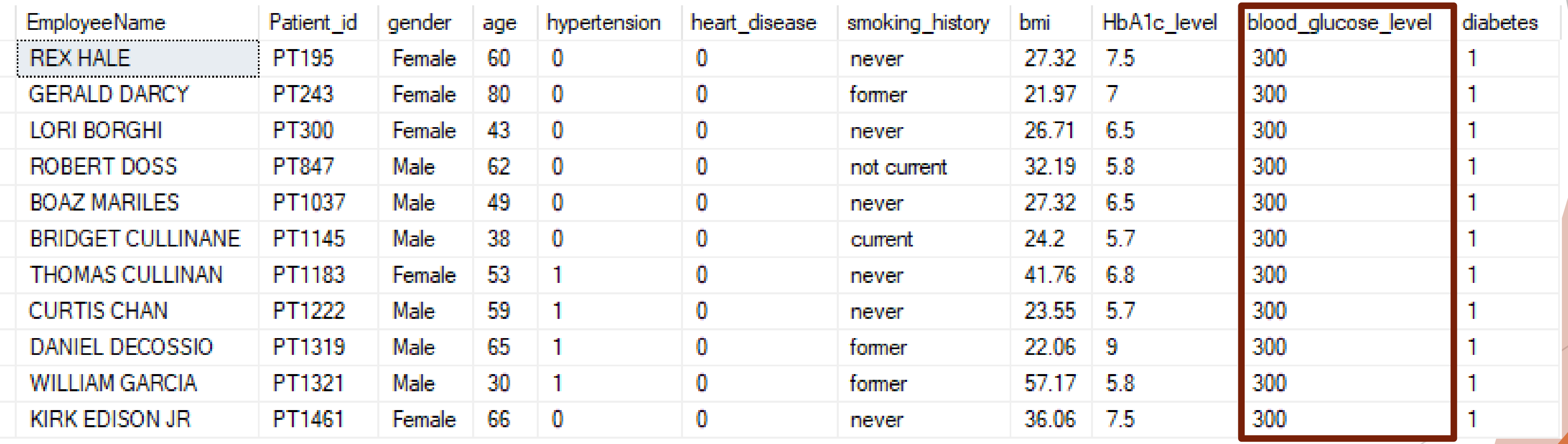


	avg_bmi
1	27.32

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

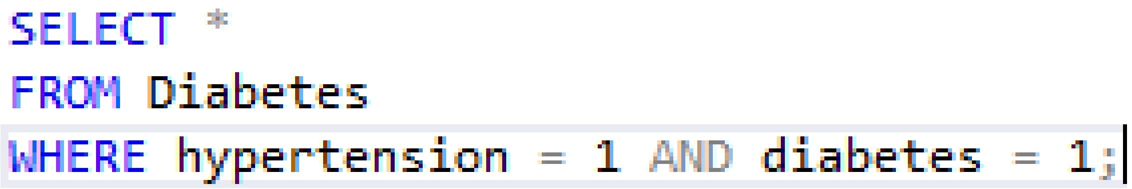


```
SELECT *
FROM Diabetes
ORDER BY blood_glucose_level DESC;
```

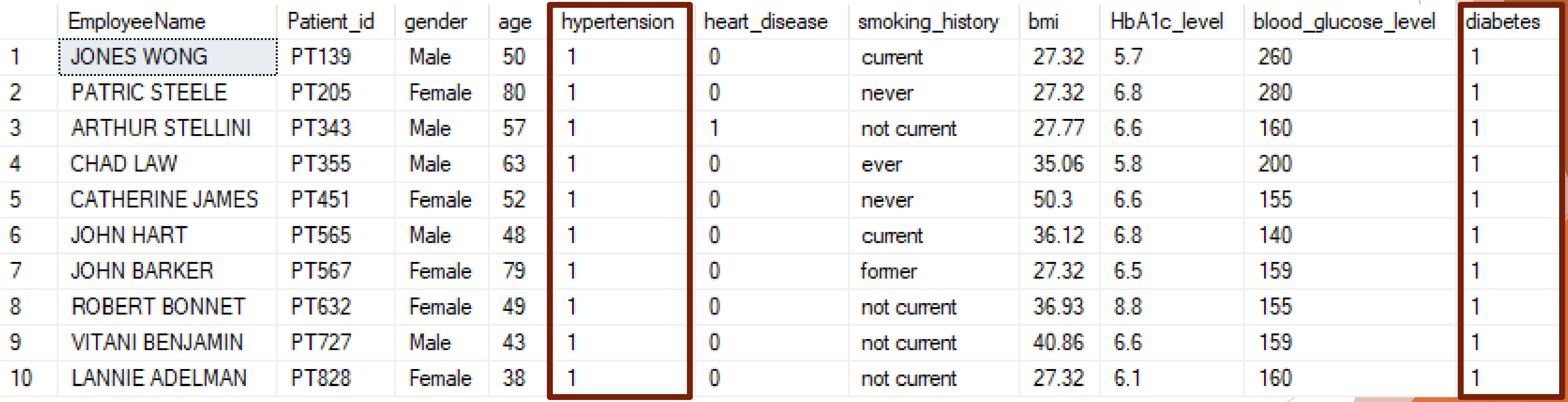


EmployeeName	Patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
REX HALE	PT195	Female	60	0	0	never	27.32	7.5	300	1
GERALD DARCY	PT243	Female	80	0	0	former	21.97	7	300	1
LORI BORGHI	PT300	Female	43	0	0	never	26.71	6.5	300	1
ROBERT DOSS	PT847	Male	62	0	0	not current	32.19	5.8	300	1
BOAZ MARILES	PT1037	Male	49	0	0	never	27.32	6.5	300	1
BRIDGET CULLINANE	PT1145	Male	38	0	0	current	24.2	5.7	300	1
THOMAS CULLINAN	PT1183	Female	53	1	0	never	41.76	6.8	300	1
CURTIS CHAN	PT1222	Male	59	1	0	never	23.55	5.7	300	1
DANIEL DECOSSIO	PT1319	Male	65	1	0	former	22.06	9	300	1
WILLIAM GARCIA	PT1321	Male	30	1	0	former	57.17	5.8	300	1
KIRK EDISON JR	PT1461	Female	66	0	0	never	36.06	7.5	300	1

5. Find patients who have hypertension and diabetes.

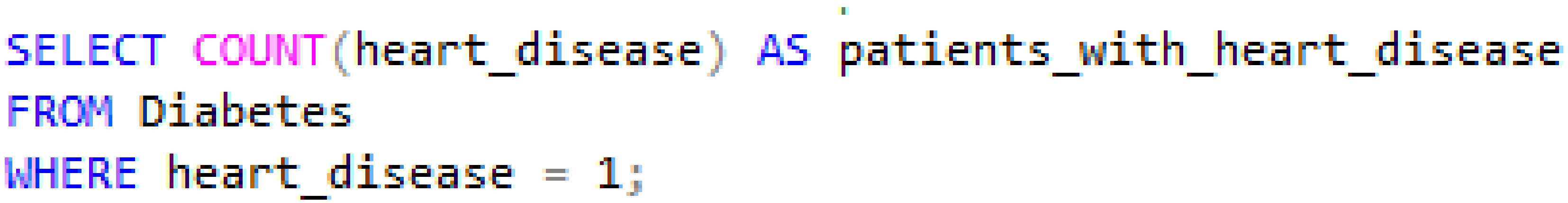


```
SELECT *
FROM Diabetes
WHERE hypertension = 1 AND diabetes = 1;
```

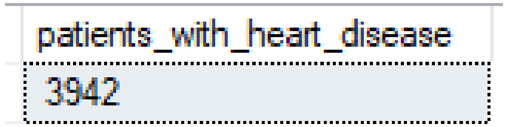


	EmployeeName	Patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
1	JONES WONG	PT139	Male	50	1	0	current	27.32	5.7	260	1
2	PATRIC STEELE	PT205	Female	80	1	0	never	27.32	6.8	280	1
3	ARTHUR STELLINI	PT343	Male	57	1	1	not current	27.77	6.6	160	1
4	CHAD LAW	PT355	Male	63	1	0	ever	35.06	5.8	200	1
5	CATHERINE JAMES	PT451	Female	52	1	0	never	50.3	6.6	155	1
6	JOHN HART	PT565	Male	48	1	0	current	36.12	6.8	140	1
7	JOHN BARKER	PT567	Female	79	1	0	former	27.32	6.5	159	1
8	ROBERT BONNET	PT632	Female	49	1	0	not current	36.93	8.8	155	1
9	VITANI BENJAMIN	PT727	Male	43	1	0	not current	40.86	6.6	159	1
10	LANNIE ADELMAN	PT828	Female	38	1	0	not current	27.32	6.1	160	1

6. Determine the number of patients with heart disease.

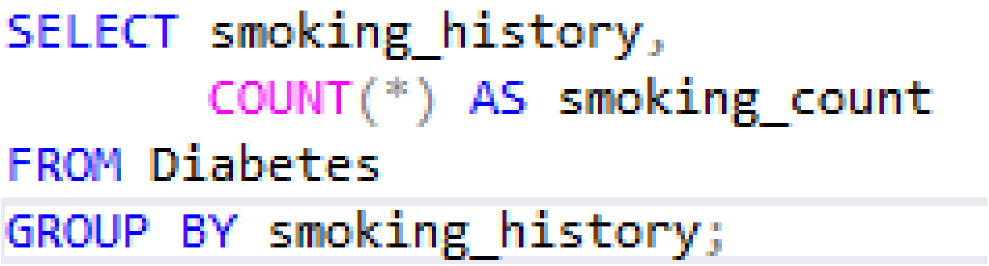


```
SELECT COUNT(heart_disease) AS patients_with_heart_disease
FROM Diabetes
WHERE heart_disease = 1;
```

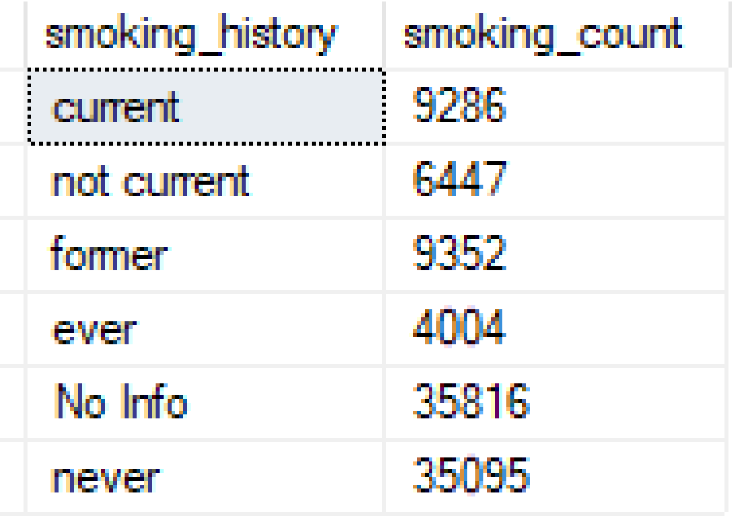


patients_with_heart_disease
3942

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

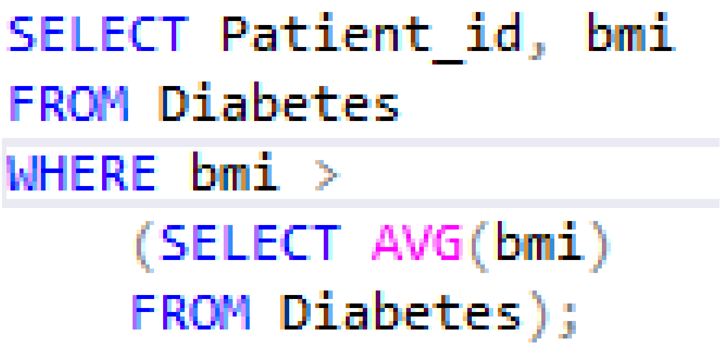


```
SELECT smoking_history,
       COUNT(*) AS smoking_count
FROM Diabetes
GROUP BY smoking_history;
```

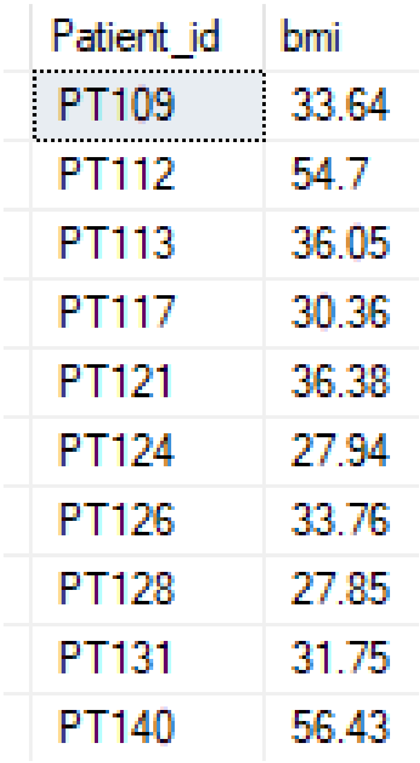


smoking_history	smoking_count
current	9286
not current	6447
former	9352
ever	4004
No Info	35816
never	35095

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

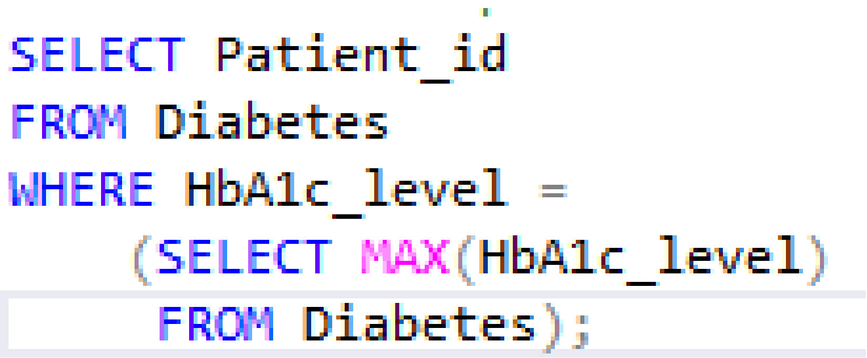


```sql
SELECT Patient_id, bmi
FROM Diabetes
WHERE bmi >
     (SELECT AVG(bmi)
     FROM Diabetes);
```

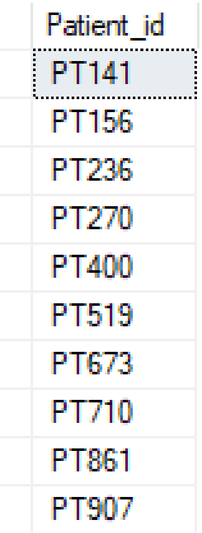


Patient_id	bmi
PT109	33.64
PT112	54.7
PT113	36.05
PT117	30.36
PT121	36.38
PT124	27.94
PT126	33.76
PT128	27.85
PT131	31.75
PT140	56.43

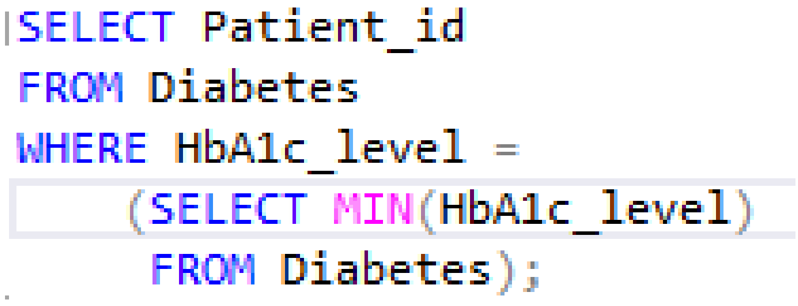
9. Find the patient with the highest HbA1c level and the patient with the lowest HbA1clevel.



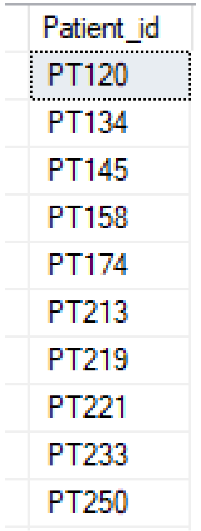
```
SELECT Patient_id
FROM Diabetes
WHERE HbA1c_level =
    (SELECT MAX(HbA1c_level)
     FROM Diabetes);
```



```
Patient_id
PT141
PT156
PT236
PT270
PT400
PT519
PT673
PT710
PT861
PT907
```

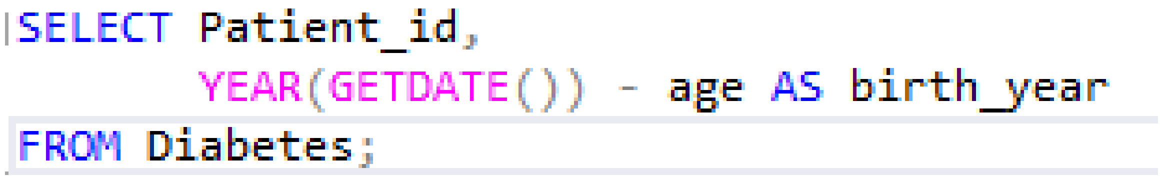


```
SELECT Patient_id
FROM Diabetes
WHERE HbA1c_level =
    (SELECT MIN(HbA1c_level)
     FROM Diabetes);
```

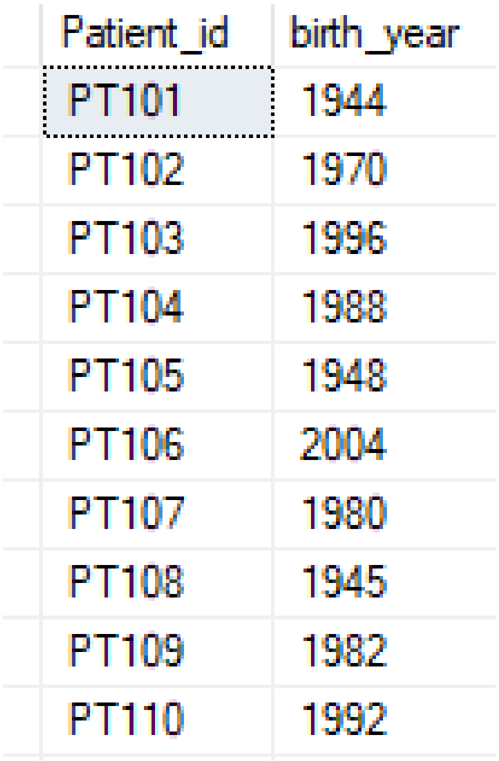


```
Patient_id
PT120
PT134
PT145
PT158
PT174
PT213
PT219
PT221
PT233
PT250
```

10. Calculate the birth year of patients (assuming the current date as of now).

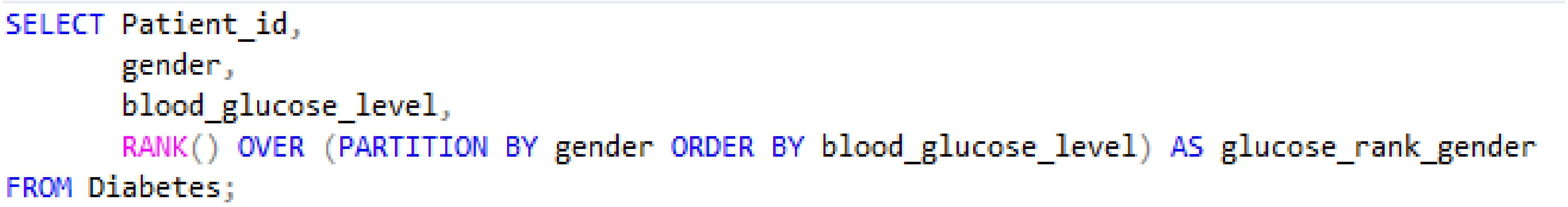


```
SELECT Patient_id,
       YEAR(GETDATE()) - age AS birth_year
FROM Diabetes;
```

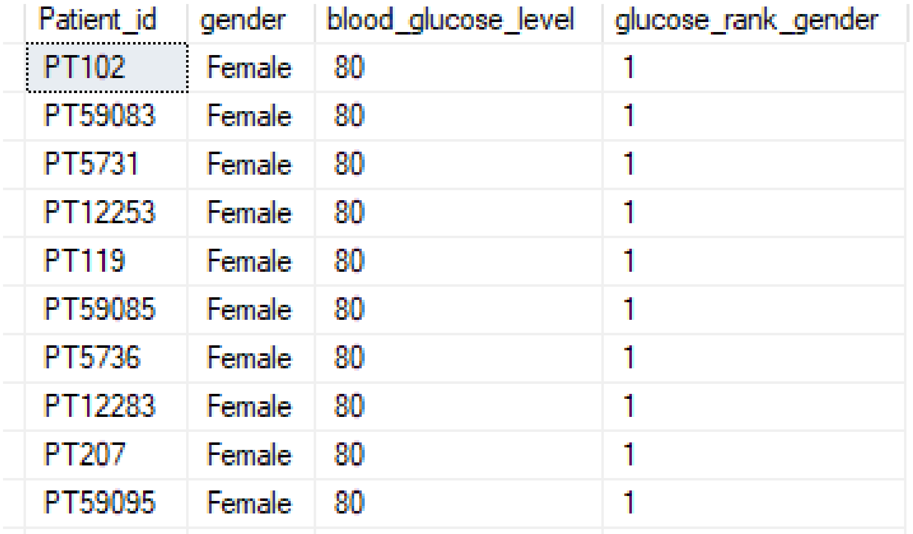


Patient_id	birth_year
PT101	1944
PT102	1970
PT103	1996
PT104	1988
PT105	1948
PT106	2004
PT107	1980
PT108	1945
PT109	1982
PT110	1992

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

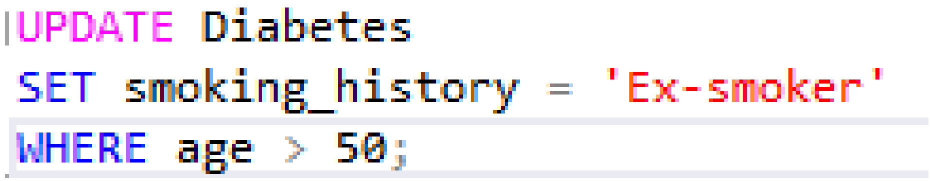


```sql
SELECT Patient_id,
       gender,
       blood_glucose_level,
       RANK() OVER (PARTITION BY gender ORDER BY blood_glucose_level) AS glucose_rank_gender
FROM Diabetes;
```

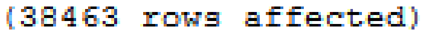


Patient_id	gender	blood_glucose_level	glucose_rank_gender
PT102	Female	80	1
PT59083	Female	80	1
PT5731	Female	80	1
PT12253	Female	80	1
PT119	Female	80	1
PT59085	Female	80	1
PT5736	Female	80	1
PT12283	Female	80	1
PT207	Female	80	1
PT59095	Female	80	1

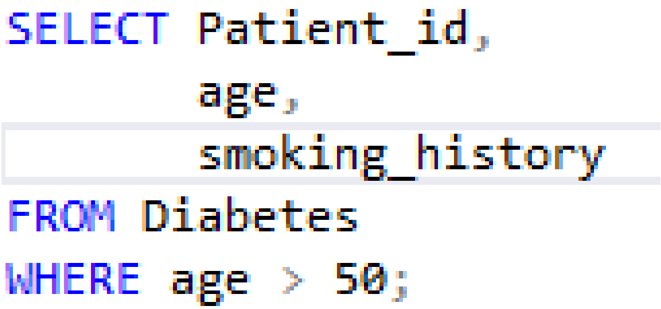
12. Update the smoking history of patients who are older than 50 to "Ex-smoker."



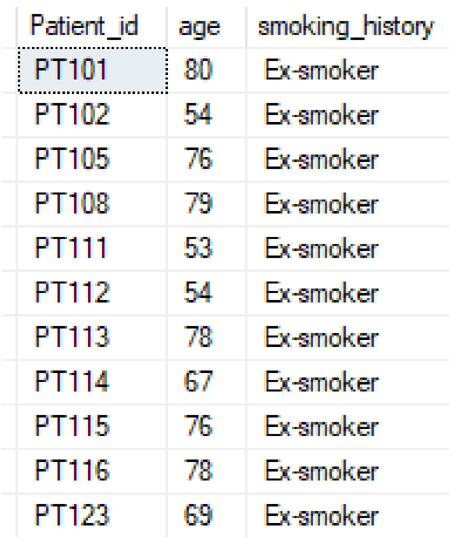
```
UPDATE Diabetes
SET smoking_history = 'Ex-smoker'
WHERE age > 50;
```



```
(38463 rows affected)
```

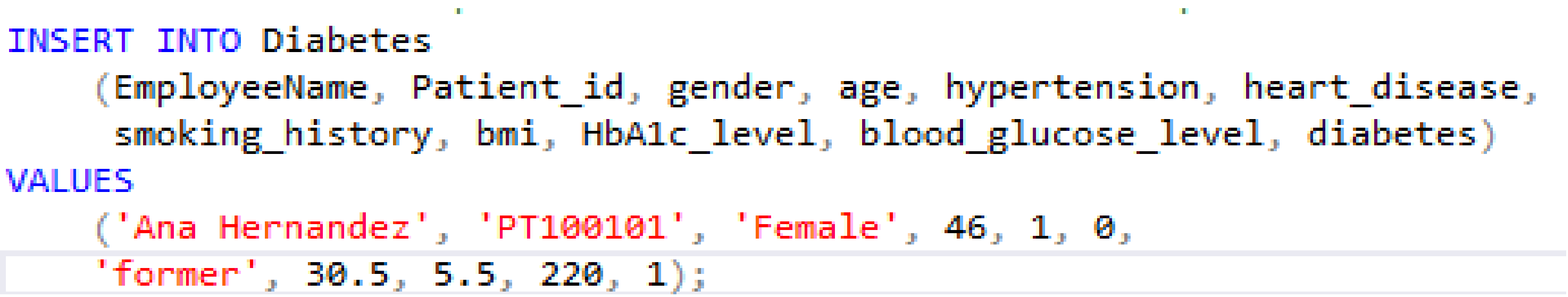


```
SELECT Patient_id,
       age,
       smoking_history
FROM Diabetes
WHERE age > 50;
```

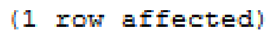


Patient_id	age	smoking_history
PT101	80	Ex-smoker
PT102	54	Ex-smoker
PT105	76	Ex-smoker
PT108	79	Ex-smoker
PT111	53	Ex-smoker
PT112	54	Ex-smoker
PT113	78	Ex-smoker
PT114	67	Ex-smoker
PT115	76	Ex-smoker
PT116	78	Ex-smoker
PT123	69	Ex-smoker

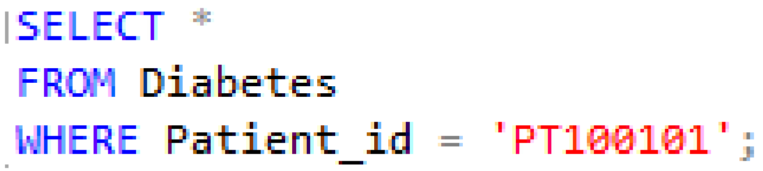
13. Insert a new patient into the database with sample data.



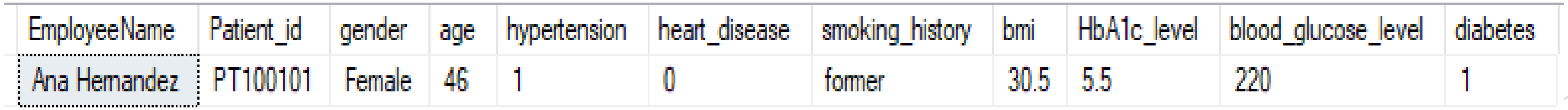
```sql
INSERT INTO Diabetes
    (EmployeeName, Patient_id, gender, age, hypertension, heart_disease,
     smoking_history, bmi, HbA1c_level, blood_glucose_level, diabetes)
VALUES
    ('Ana Hernandez', 'PT100101', 'Female', 46, 1, 0,
    'former', 30.5, 5.5, 220, 1);
```



```
(1 row affected)
```

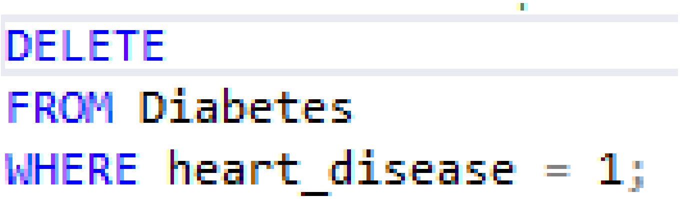


```sql
SELECT *
FROM Diabetes
WHERE Patient_id = 'PT100101';
```

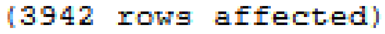


EmployeeName	Patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Ana Hernandez	PT100101	Female	46	1	0	former	30.5	5.5	220	1

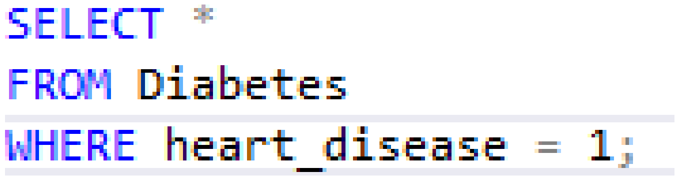
14. Delete all patients with heart disease from the database.



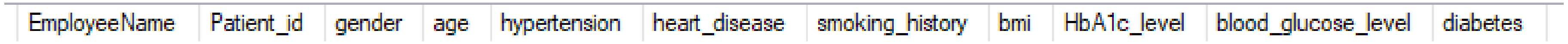
```sql
DELETE
FROM Diabetes
WHERE heart_disease = 1;
```



```
(3942 rows affected)
```

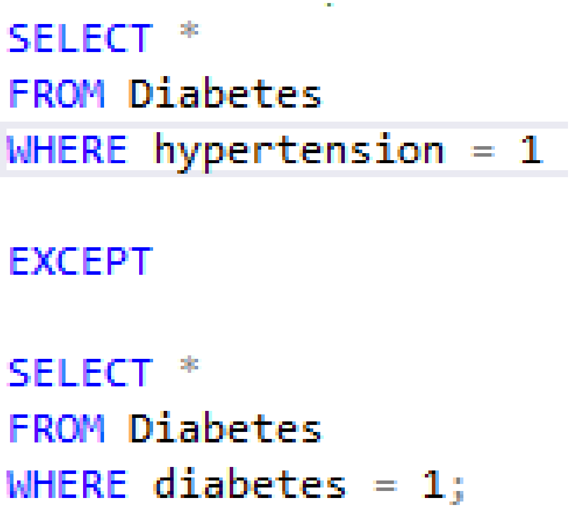


```sql
SELECT *
FROM Diabetes
WHERE heart_disease = 1;
```

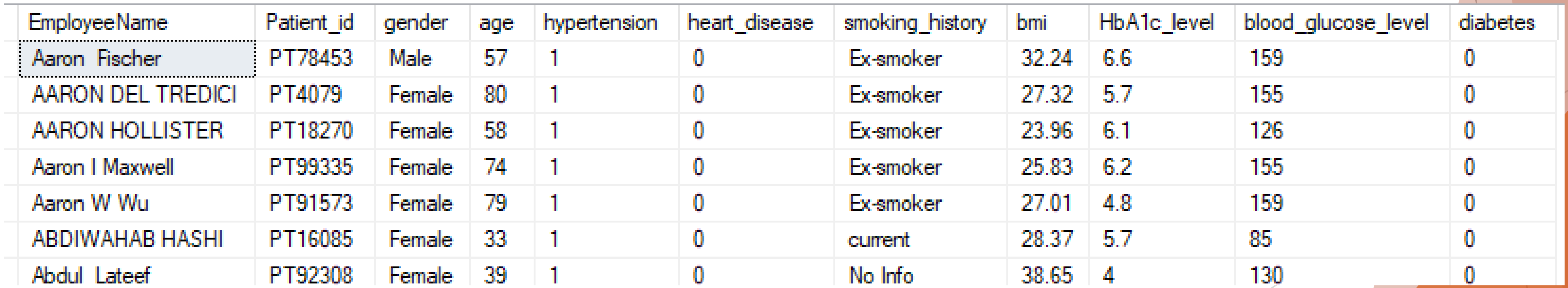


EmployeeName	Patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes

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

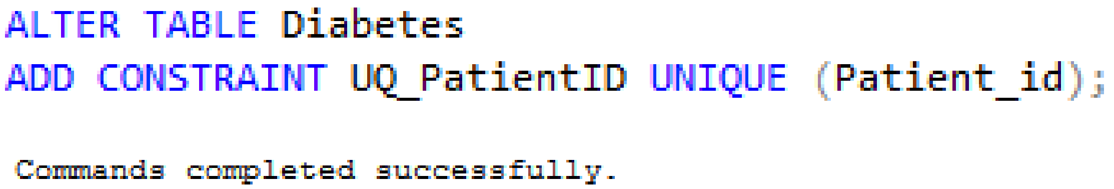


```
SELECT *
FROM Diabetes
WHERE hypertension = 1

EXCEPT

SELECT *
FROM Diabetes
WHERE diabetes = 1;
```



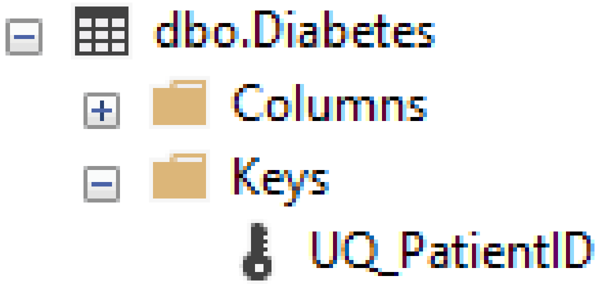
EmployeeName	Patient_id	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Aaron Fischer	PT78453	Male	57	1	0	Ex-smoker	32.24	6.6	159	0
AARON DEL TREDICI	PT4079	Female	80	1	0	Ex-smoker	27.32	5.7	155	0
AARON HOLLISTER	PT18270	Female	58	1	0	Ex-smoker	23.96	6.1	126	0
Aaron I Maxwell	PT99335	Female	74	1	0	Ex-smoker	25.83	6.2	155	0
Aaron W Wu	PT91573	Female	79	1	0	Ex-smoker	27.01	4.8	159	0
ABDIWAHAB HASHI	PT16085	Female	33	1	0	current	28.37	5.7	85	0
Abdul Lateef	PT92308	Female	39	1	0	No Info	38.65	4	130	0

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

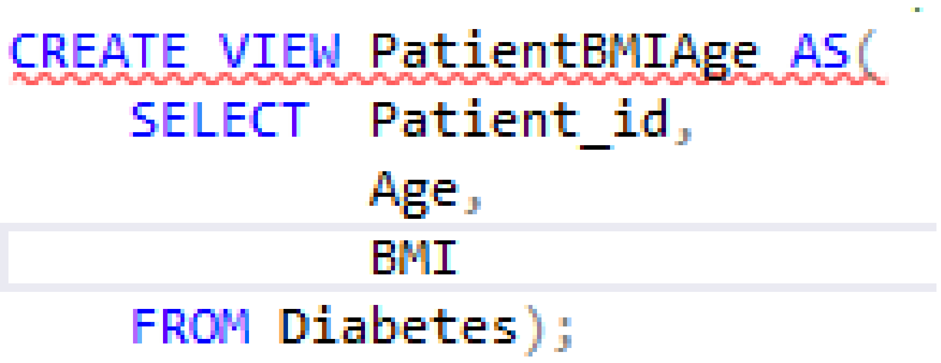


```
ALTER TABLE Diabetes
ADD CONSTRAINT UQ_PatientID UNIQUE (Patient_id);
```

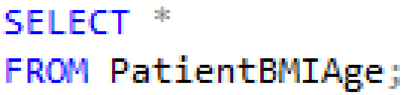
```
Commands completed successfully.
```



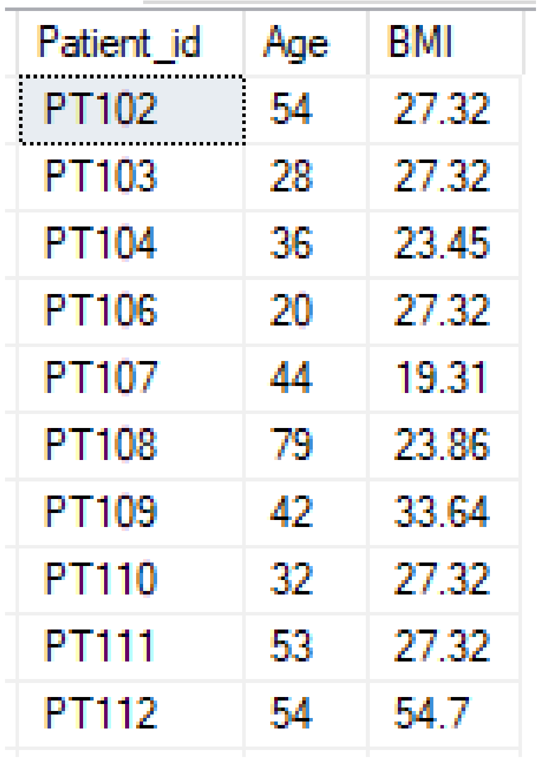
17. Create a view that displays the Patient_ids, ages, and BMI of patients



```
CREATE VIEW PatientBMIAge AS(
    SELECT  Patient_id,
            Age,
            BMI
    FROM Diabetes);
```



```
SELECT *
FROM PatientBMIAge;
```



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
PT112	54	54.7

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

- *Normalize tables*: create a table for patient’s demographics and other for health parameters, adding the date to each record to track the patient’s health status.
- *Use primary and foreign keys* to relate each table and avoid duplicates.
- *Use appropiate data types* to make sure that the new data registered is consistent to the requested information.
- *Apply constraints* like NOT NULL to the columns that must be filled.
- Delete redundant columns: analyze the information in each column to decide if is really important to keep or if we can obtain similar information from another variable.

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

- Select only the columns needed, and not all of them.
- Use efficiently the WHERE clause to work with only the necessary records.
- Prefer the use of joins instead of subqueries whenever it’s posible.
- Use stored procedures when creating complex instructions.
- Use windows to work with the data of interest.
- Create indexes in columns frequently used inside the WHERE clause or joins.