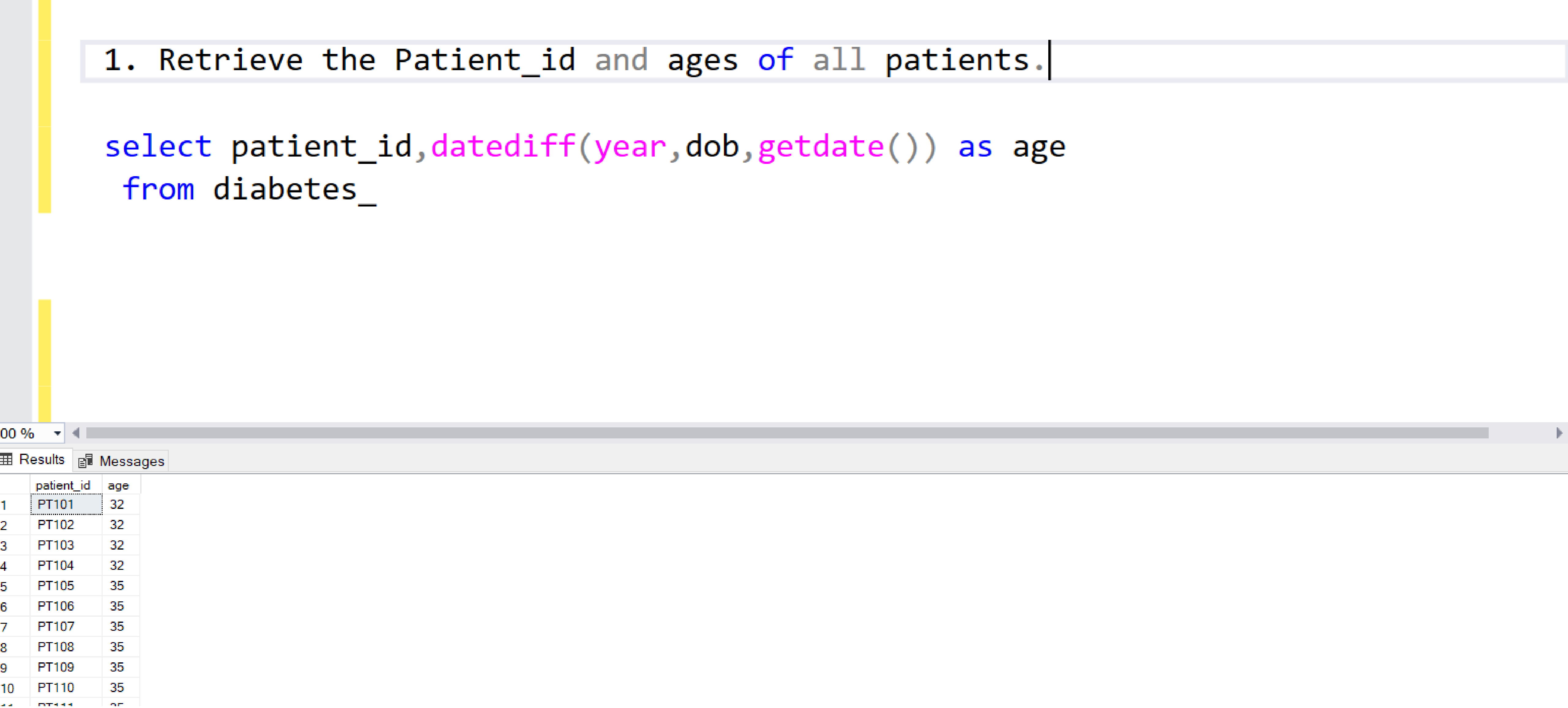
DIABETES PREDICATION PROJECT

Overview

In this project , we will leverage SQL querying and data analysis skills to analyse comprehensive dataset containing demographic, clinical and lifestyle information of individuals.

The dataset will include variables such as Patient Name, age, gender, body mass index (BMI),blood pressure, heart disease, smoking history, blood glucose level, patient is diabetic.

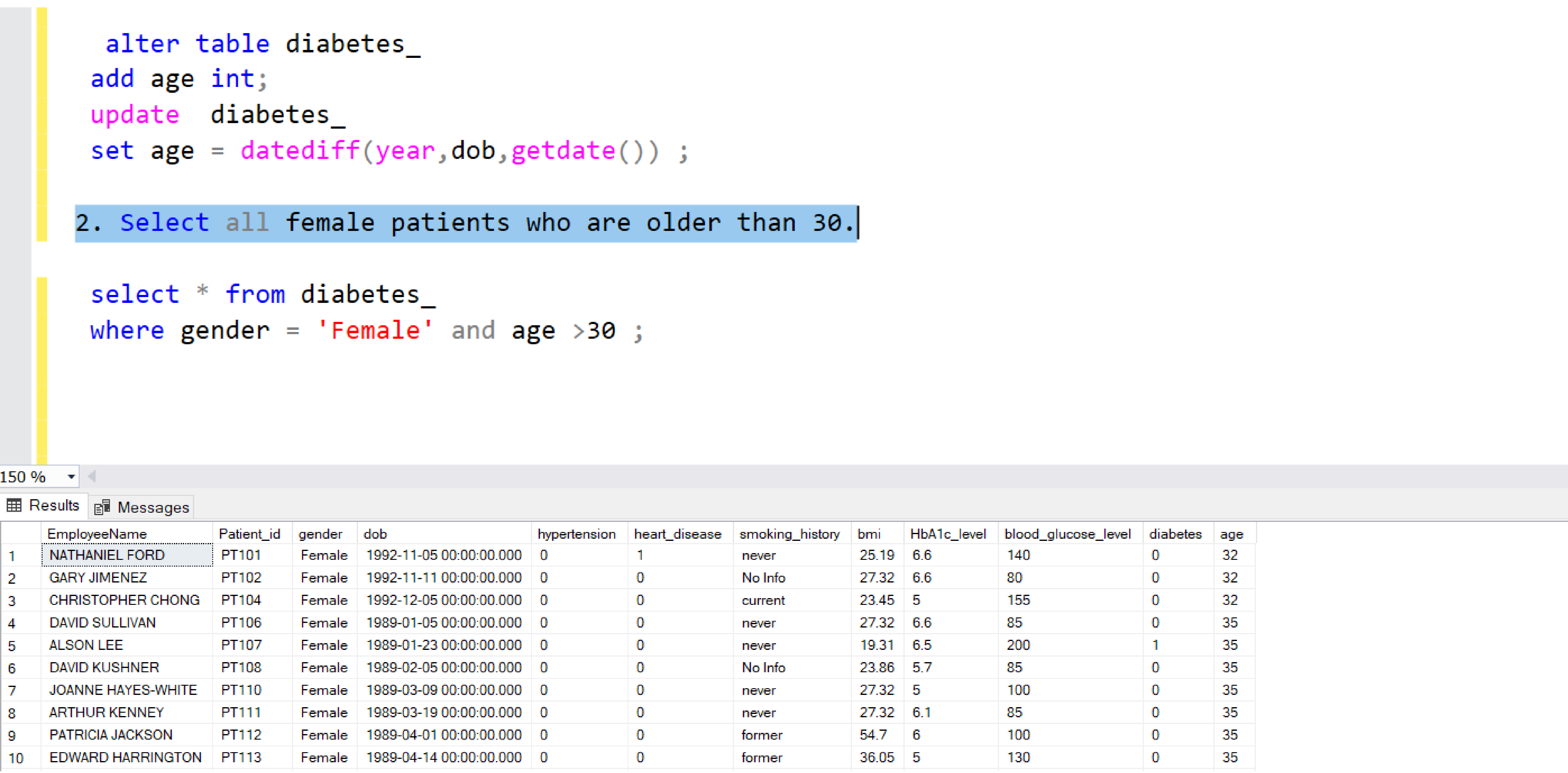
AAKASH SHARMA



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

select patient_id,datediff(year,dob,getdate()) as age
 from diabetes_
```

	patient_id	age
1	PT101	32
2	PT102	32
3	PT103	32
4	PT104	32
5	PT105	35
6	PT106	35
7	PT107	35
8	PT108	35
9	PT109	35
10	PT110	35
11	PT111	35

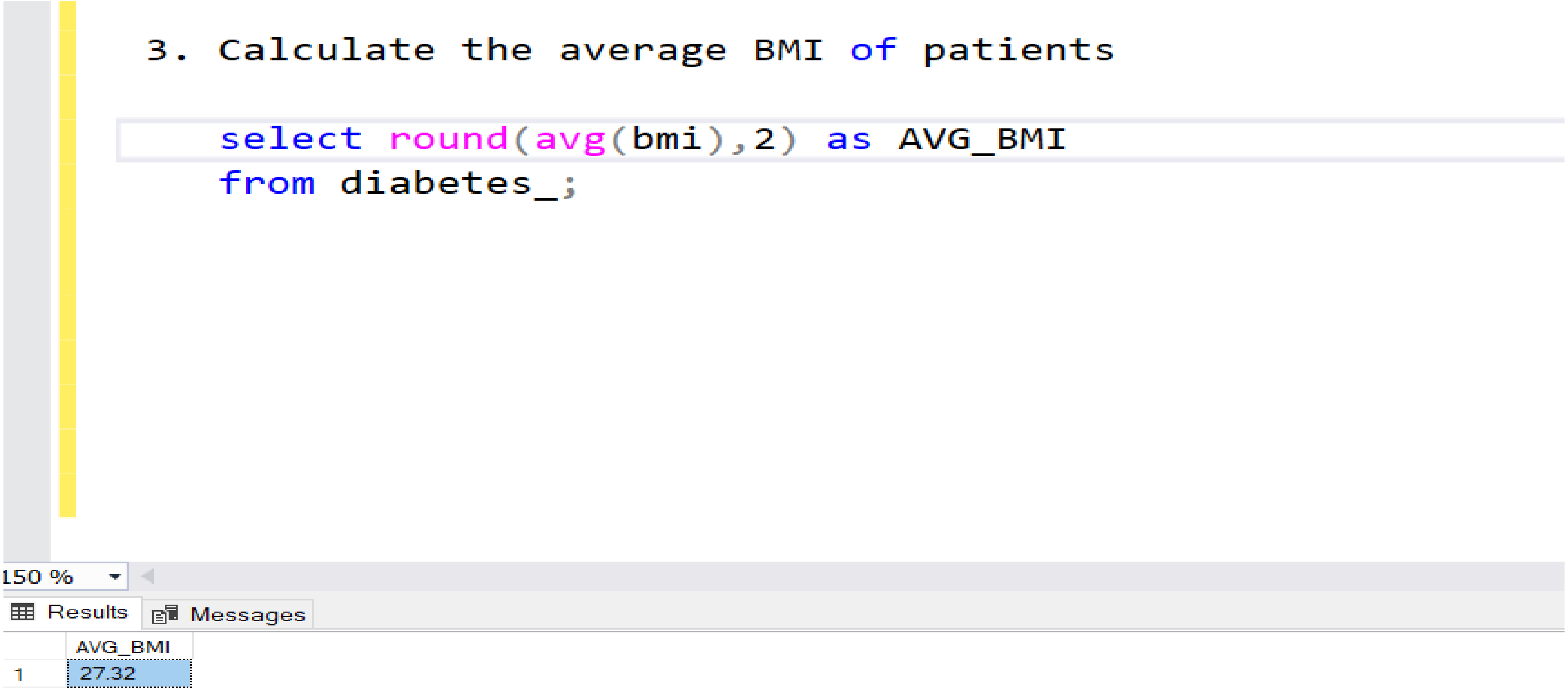


```
 alter table diabetes_
add age int;
update  diabetes_
set age = datediff(year,dob,getdate()) ;

2. Select all female patients who are older than 30.

select * from diabetes_
where gender = 'Female' and age >30 ;
```

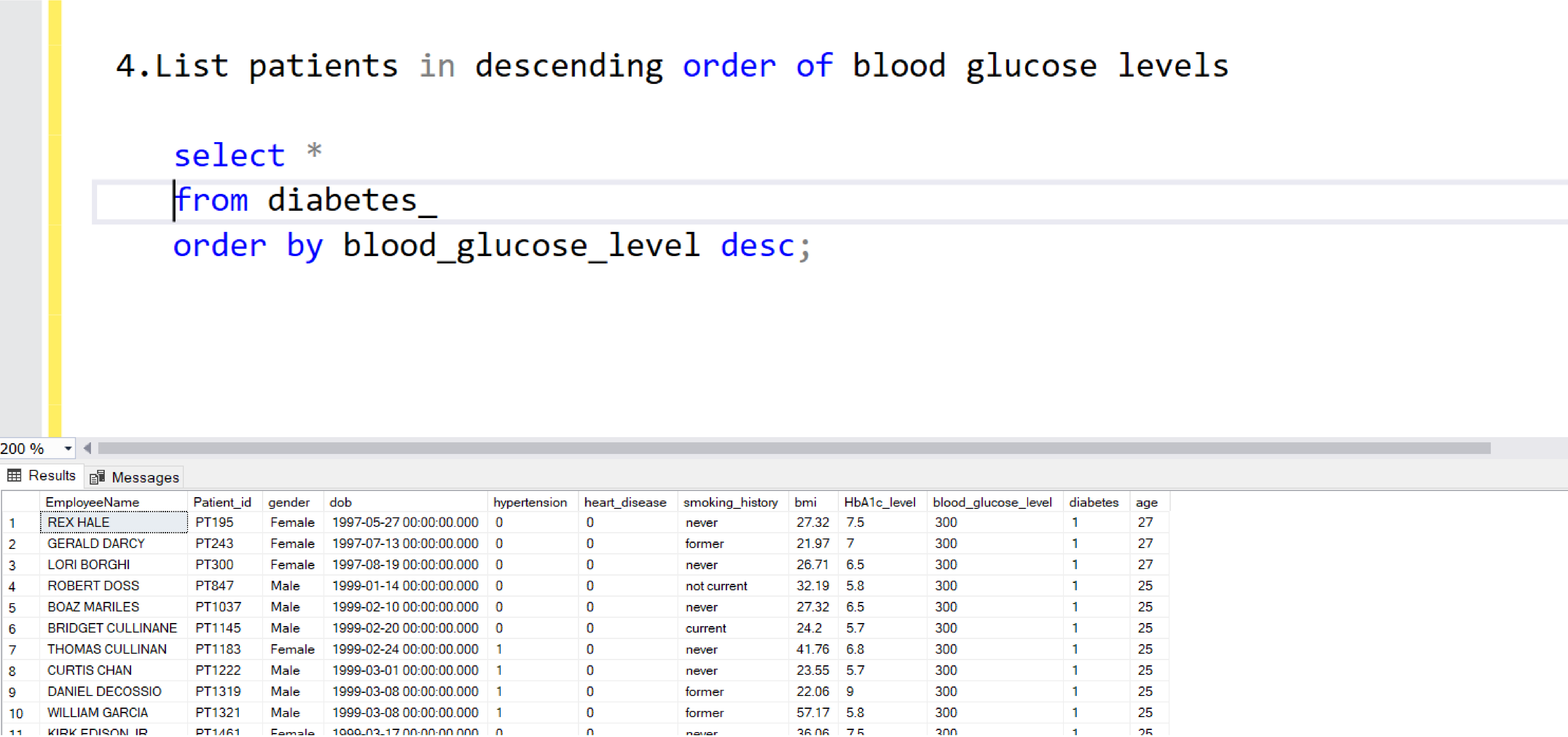
	EmployeeName	Patient_id	gender	dob	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes	age
1	NATHANIEL FORD	PT101	Female	1992-11-05 00:00:00.000	0	1	never	25.19	6.6	140	0	32
2	GARY JIMENEZ	PT102	Female	1992-11-11 00:00:00.000	0	0	No Info	27.32	6.6	80	0	32
3	CHRISTOPHER CHONG	PT104	Female	1992-12-05 00:00:00.000	0	0	current	23.45	5	155	0	32
4	DAVID SULLIVAN	PT106	Female	1989-01-05 00:00:00.000	0	0	never	27.32	6.6	85	0	35
5	ALSON LEE	PT107	Female	1989-01-23 00:00:00.000	0	0	never	19.31	6.5	200	1	35
6	DAVID KUSHNER	PT108	Female	1989-02-05 00:00:00.000	0	0	No Info	23.86	5.7	85	0	35
7	JOANNE HAYES-WHITE	PT110	Female	1989-03-09 00:00:00.000	0	0	never	27.32	5	100	0	35
8	ARTHUR KENNEY	PT111	Female	1989-03-19 00:00:00.000	0	0	never	27.32	6.1	85	0	35
9	PATRICIA JACKSON	PT112	Female	1989-04-01 00:00:00.000	0	0	former	54.7	6	100	0	35
10	EDWARD HARRINGTON	PT113	Female	1989-04-14 00:00:00.000	0	0	former	36.05	5	130	0	35



```
3. Calculate the average BMI of patients

    select round(avg(bmi),2) as AVG_BMI
    from diabetes_;
```

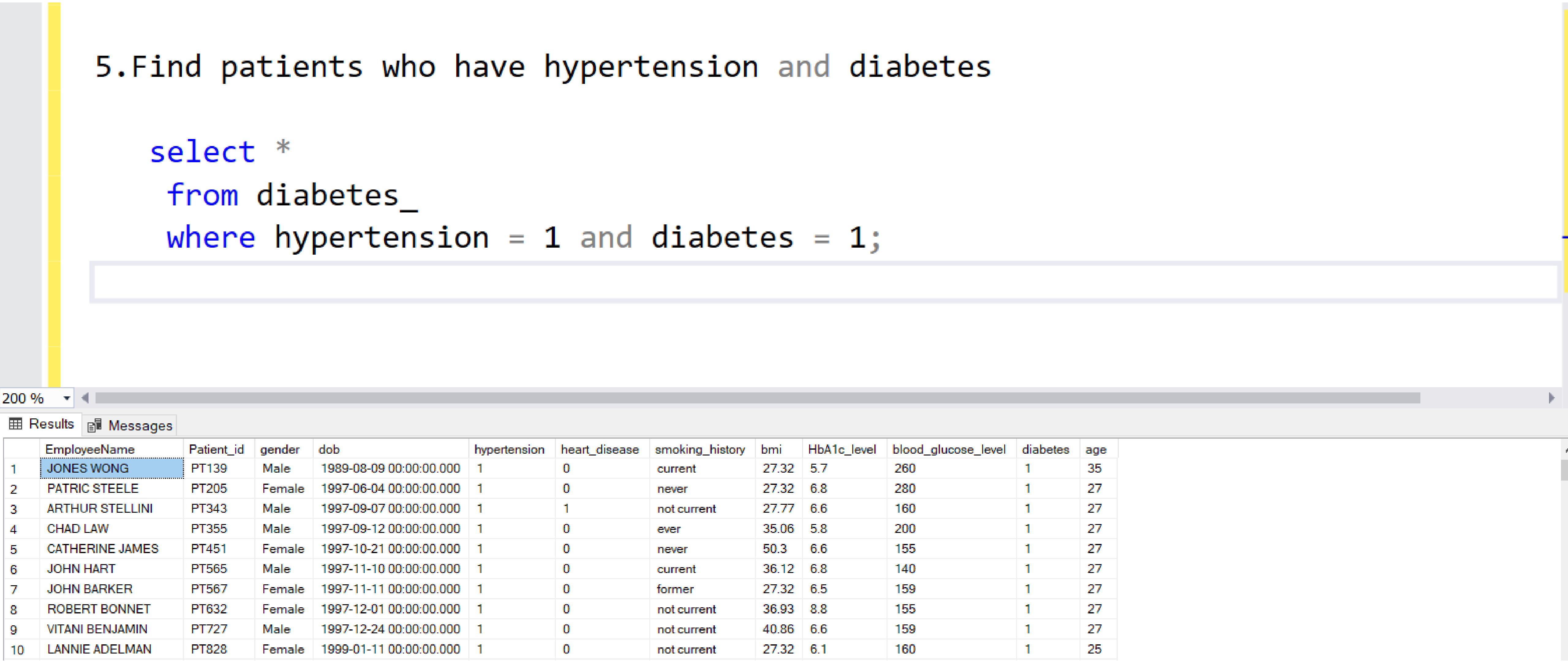
```
     AVG_BMI
1    27.32
```



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

    select *
    from diabetes_
    order by blood_glucose_level desc;
```

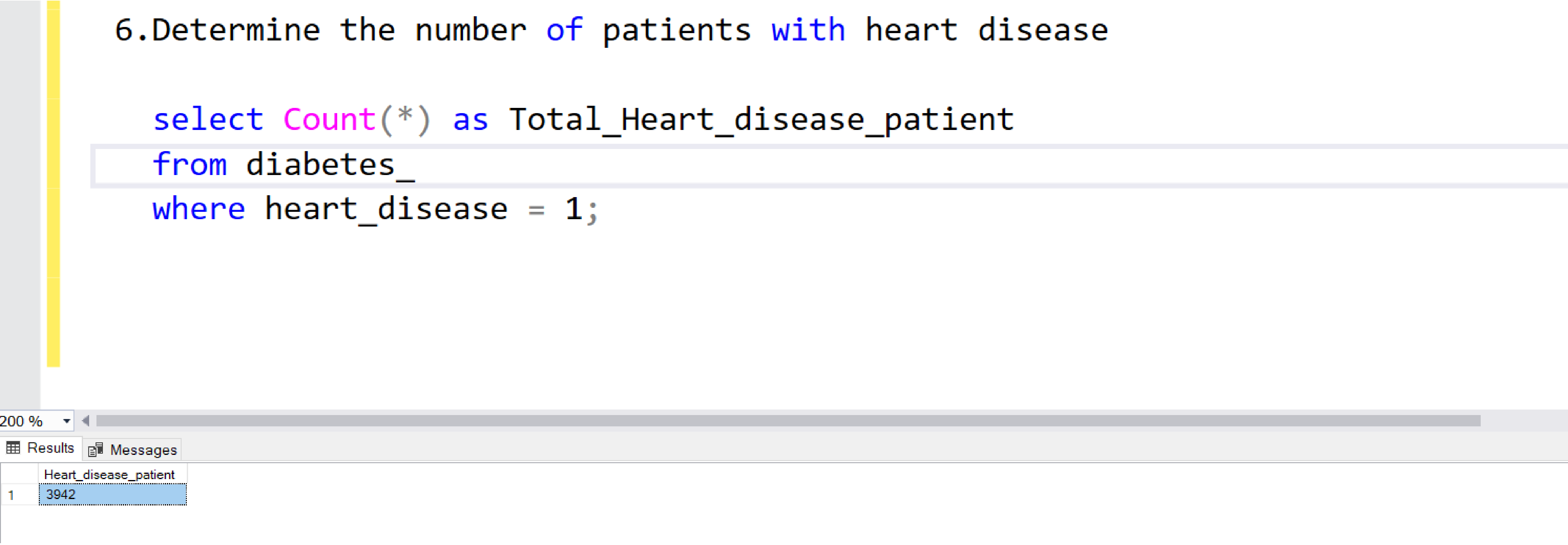
	EmployeeName	Patient_id	gender	dob	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes	age
1	REX HALE	PT195	Female	1997-05-27 00:00:00.000	0	0	never	27.32	7.5	300	1	27
2	GERALD DARCY	PT243	Female	1997-07-13 00:00:00.000	0	0	former	21.97	7	300	1	27
3	LORI BORGHI	PT300	Female	1997-08-19 00:00:00.000	0	0	never	26.71	6.5	300	1	27
4	ROBERT DOSS	PT847	Male	1999-01-14 00:00:00.000	0	0	not current	32.19	5.8	300	1	25
5	BOAZ MARILES	PT1037	Male	1999-02-10 00:00:00.000	0	0	never	27.32	6.5	300	1	25
6	BRIDGET CULLINANE	PT1145	Male	1999-02-20 00:00:00.000	0	0	current	24.2	5.7	300	1	25
7	THOMAS CULLINAN	PT1183	Female	1999-02-24 00:00:00.000	1	0	never	41.76	6.8	300	1	25
8	CURTIS CHAN	PT1222	Male	1999-03-01 00:00:00.000	1	0	never	23.55	5.7	300	1	25
9	DANIEL DECOSSIO	PT1319	Male	1999-03-08 00:00:00.000	1	0	former	22.06	9	300	1	25
10	WILLIAM GARCIA	PT1321	Male	1999-03-08 00:00:00.000	1	0	former	57.17	5.8	300	1	25
11	KIRK EDISON JR	PT1461	Female	1999-03-17 00:00:00.000	0	0	never	36.06	7.5	300	1	25



```
5.Find patients who have hypertension and diabetes

    select *
     from diabetes_
     where hypertension = 1 and diabetes = 1;
```

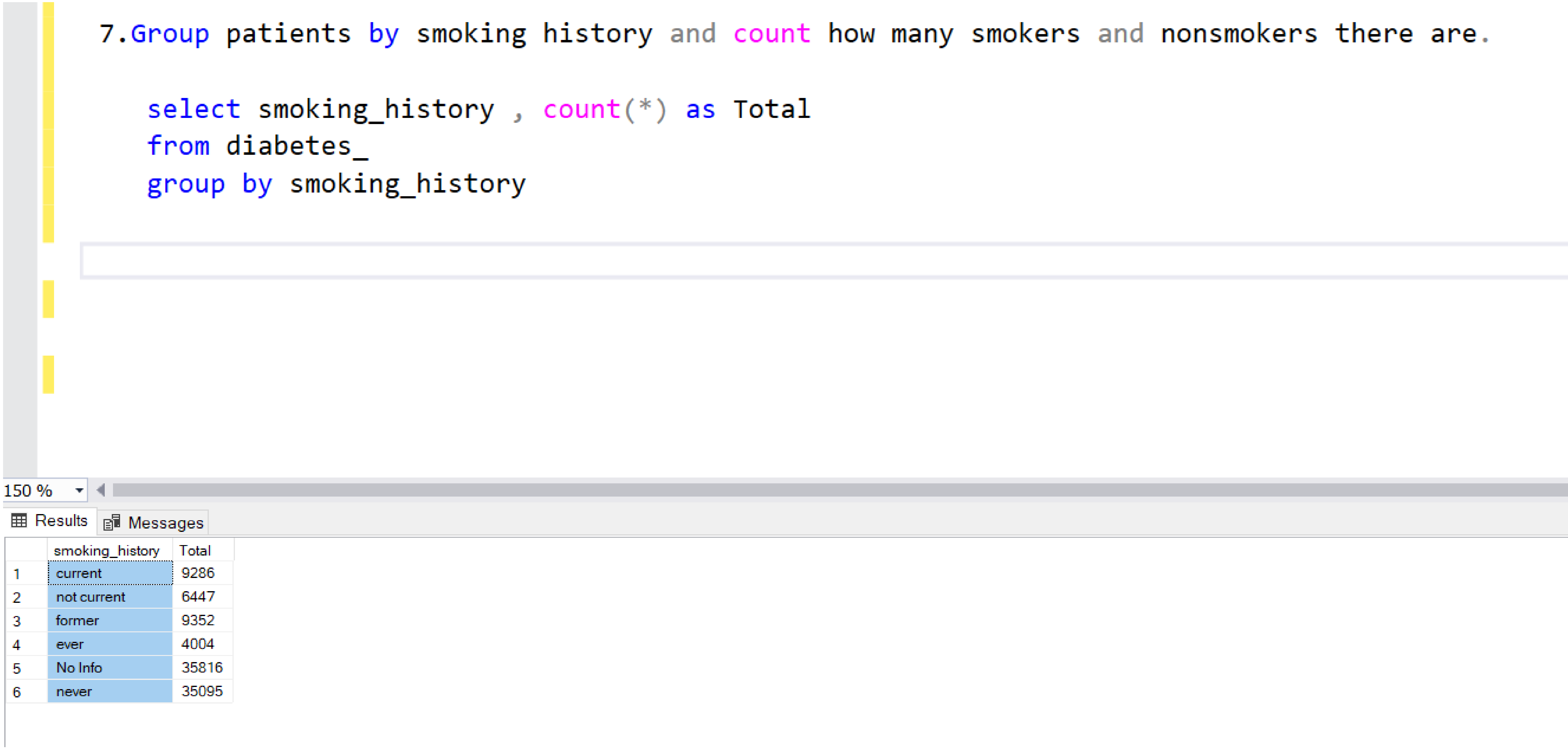
	EmployeeName	Patient_id	gender	dob	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes	age
1	JONES WONG	PT139	Male	1989-08-09 00:00:00.000	1	0	current	27.32	5.7	260	1	35
2	PATRIC STEELE	PT205	Female	1997-06-04 00:00:00.000	1	0	never	27.32	6.8	280	1	27
3	ARTHUR STELLINI	PT343	Male	1997-09-07 00:00:00.000	1	1	not current	27.77	6.6	160	1	27
4	CHAD LAW	PT355	Male	1997-09-12 00:00:00.000	1	0	ever	35.06	5.8	200	1	27
5	CATHERINE JAMES	PT451	Female	1997-10-21 00:00:00.000	1	0	never	50.3	6.6	155	1	27
6	JOHN HART	PT565	Male	1997-11-10 00:00:00.000	1	0	current	36.12	6.8	140	1	27
7	JOHN BARKER	PT567	Female	1997-11-11 00:00:00.000	1	0	former	27.32	6.5	159	1	27
8	ROBERT BONNET	PT632	Female	1997-12-01 00:00:00.000	1	0	not current	36.93	8.8	155	1	27
9	VITANI BENJAMIN	PT727	Male	1997-12-24 00:00:00.000	1	0	not current	40.86	6.6	159	1	27
10	LANNIE ADELMAN	PT828	Female	1999-01-11 00:00:00.000	1	0	not current	27.32	6.1	160	1	25



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

  select Count(*) as Total_Heart_disease_patient
  from diabetes_
  where heart_disease = 1;
```

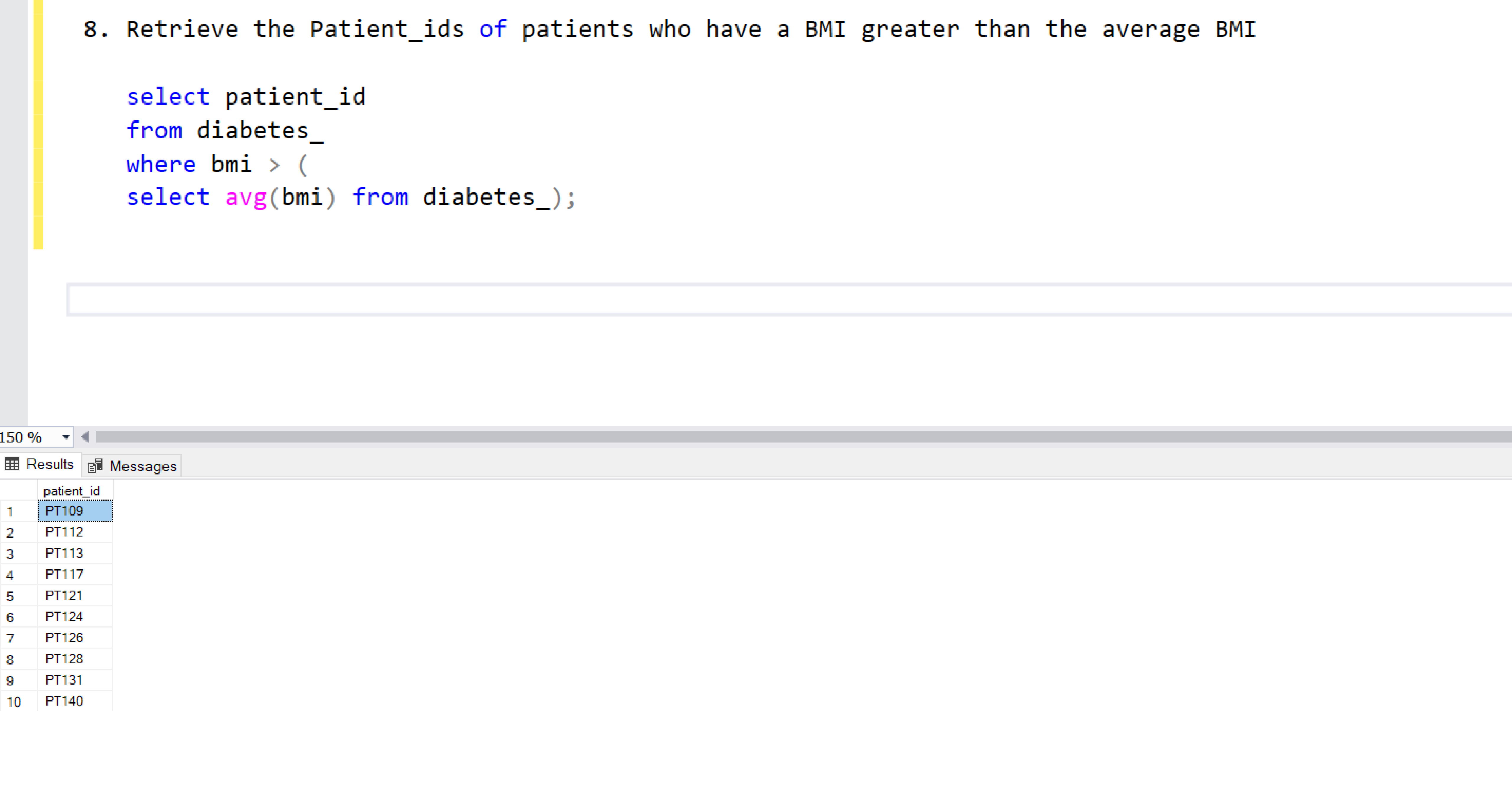
```
   Heart_disease_patient
1  3942
```



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

   select smoking_history , count(*) as Total
   from diabetes_
   group by smoking_history
```

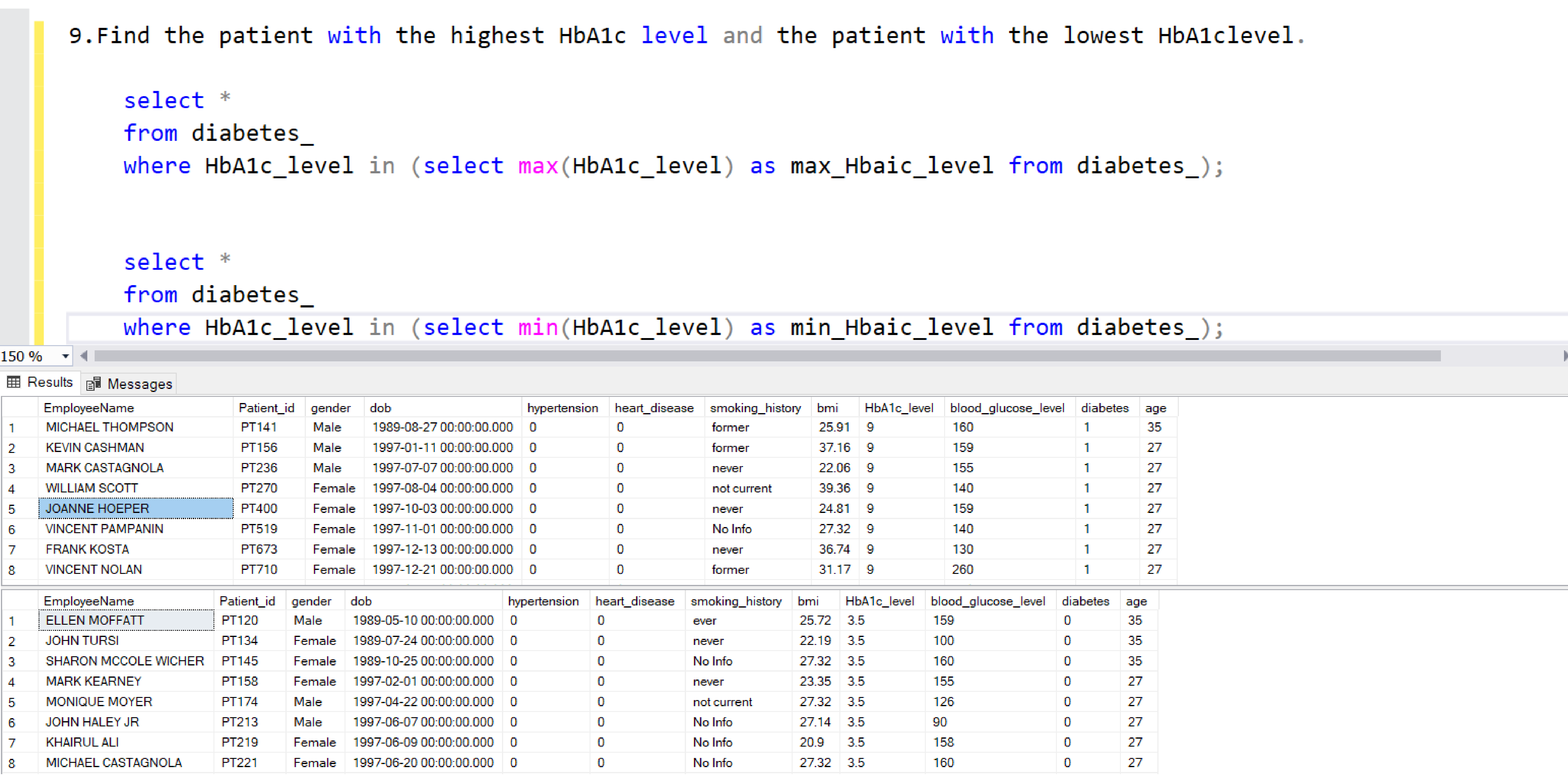
	smoking_history	Total
1	current	9286
2	not current	6447
3	former	9352
4	ever	4004
5	No Info	35816
6	never	35095



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

   select patient_id
   from diabetes_
   where bmi > (
   select avg(bmi) from diabetes_);
```

	patient_id
1	PT109
2	PT112
3	PT113
4	PT117
5	PT121
6	PT124
7	PT126
8	PT128
9	PT131
10	PT140



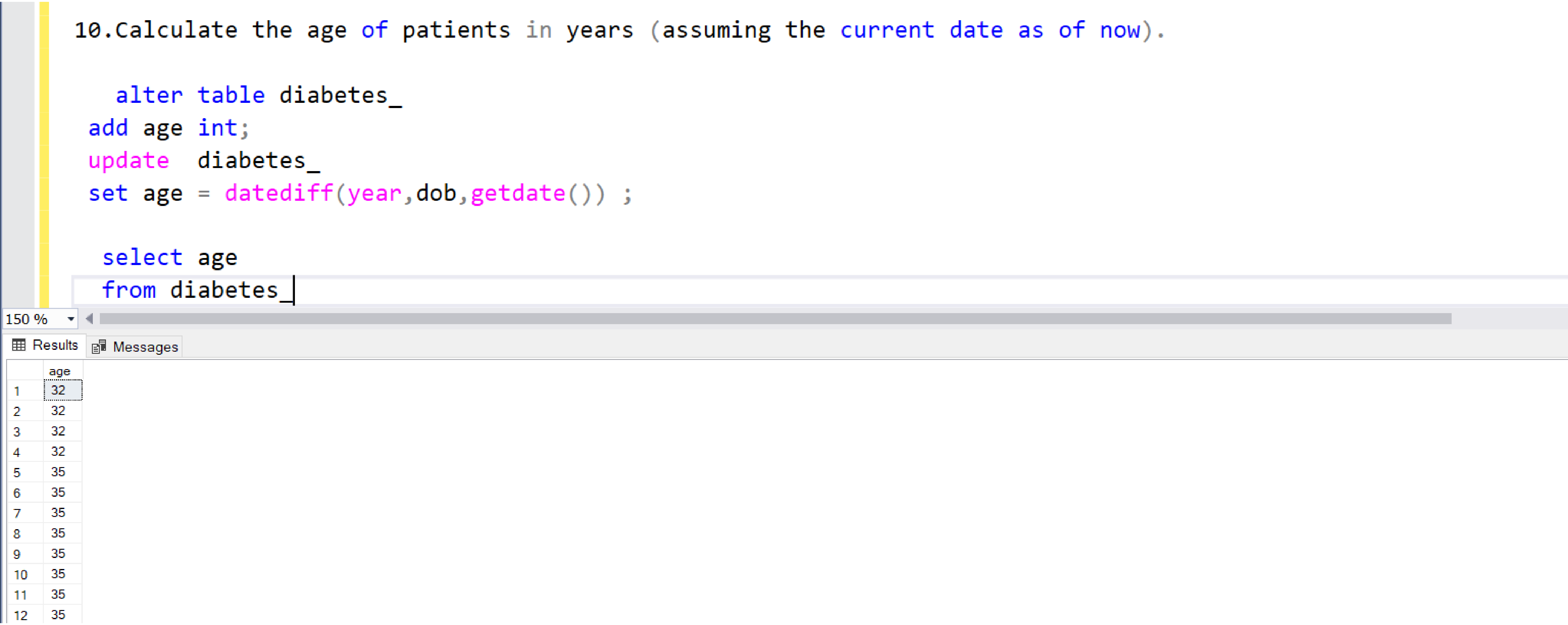
```
9.Find the patient with the highest HbA1c level and the patient with the lowest HbA1clevel.

    select *
    from diabetes_
    where HbA1c_level in (select max(HbA1c_level) as max_Hbaic_level from diabetes_);


    select *
    from diabetes_
    where HbA1c_level in (select min(HbA1c_level) as min_Hbaic_level from diabetes_);
```

	EmployeeName	Patient_id	gender	dob	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes	age
1	MICHAEL THOMPSON	PT141	Male	1989-08-27 00:00:00.000	0	0	former	25.91	9	160	1	35
2	KEVIN CASHMAN	PT156	Male	1997-01-11 00:00:00.000	0	0	former	37.16	9	159	1	27
3	MARK CASTAGNOLA	PT236	Male	1997-07-07 00:00:00.000	0	0	never	22.06	9	155	1	27
4	WILLIAM SCOTT	PT270	Female	1997-08-04 00:00:00.000	0	0	not current	39.36	9	140	1	27
5	JOANNE HOEPER	PT400	Female	1997-10-03 00:00:00.000	0	0	never	24.81	9	159	1	27
6	VINCENT PAMPANIN	PT519	Female	1997-11-01 00:00:00.000	0	0	No Info	27.32	9	140	1	27
7	FRANK KOSTA	PT673	Female	1997-12-13 00:00:00.000	0	0	never	36.74	9	130	1	27
8	VINCENT NOLAN	PT710	Female	1997-12-21 00:00:00.000	0	0	former	31.17	9	260	1	27

	EmployeeName	Patient_id	gender	dob	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes	age
1	ELLEN MOFFATT	PT120	Male	1989-05-10 00:00:00.000	0	0	ever	25.72	3.5	159	0	35
2	JOHN TURSI	PT134	Female	1989-07-24 00:00:00.000	0	0	never	22.19	3.5	100	0	35
3	SHARON MCCOLE WICHER	PT145	Female	1989-10-25 00:00:00.000	0	0	No Info	27.32	3.5	160	0	35
4	MARK KEARNEY	PT158	Female	1997-02-01 00:00:00.000	0	0	never	23.35	3.5	155	0	27
5	MONIQUE MOYER	PT174	Male	1997-04-22 00:00:00.000	0	0	not current	27.32	3.5	126	0	27
6	JOHN HALEY JR	PT213	Male	1997-06-07 00:00:00.000	0	0	No Info	27.14	3.5	90	0	27
7	KHAIRUL ALI	PT219	Female	1997-06-09 00:00:00.000	0	0	No Info	20.9	3.5	158	0	27
8	MICHAEL CASTAGNOLA	PT221	Female	1997-06-20 00:00:00.000	0	0	No Info	27.32	3.5	160	0	27

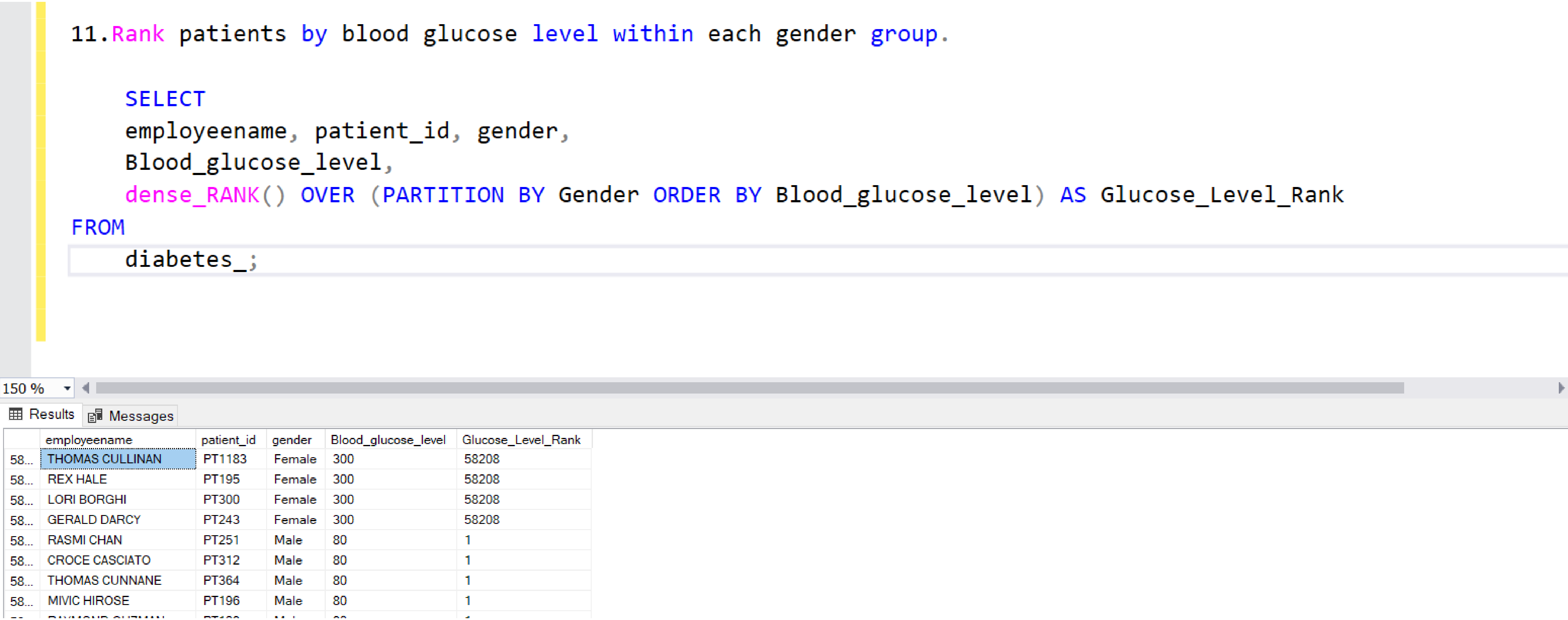


```
10.Calculate the age of patients in years (assuming the current date as of now).

   alter table diabetes_
 add age int;
 update  diabetes_
 set age = datediff(year,dob,getdate()) ;

  select age
  from diabetes_
```

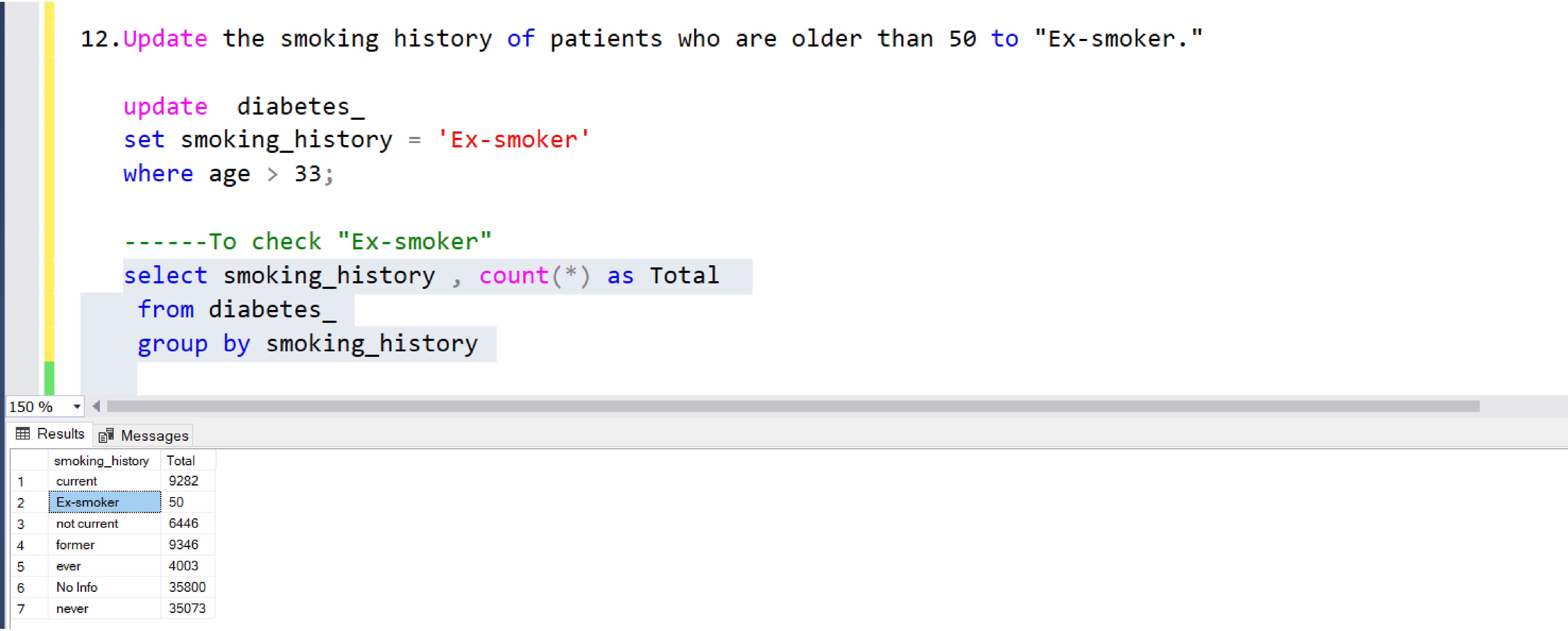
	age
1	32
2	32
3	32
4	32
5	35
6	35
7	35
8	35
9	35
10	35
11	35
12	35



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

    SELECT
    employeename, patient_id, gender,
    Blood_glucose_level,
    dense_RANK() OVER (PARTITION BY Gender ORDER BY Blood_glucose_level) AS Glucose_Level_Rank
FROM
    diabetes_;
```

	employeename	patient_id	gender	Blood_glucose_level	Glucose_Level_Rank
58...	THOMAS CULLINAN	PT1183	Female	300	58208
58...	REX HALE	PT195	Female	300	58208
58...	LORI BORGHI	PT300	Female	300	58208
58...	GERALD DARCY	PT243	Female	300	58208
58...	RASMI CHAN	PT251	Male	80	1
58...	CROCE CASCIATO	PT312	Male	80	1
58...	THOMAS CUNNANE	PT364	Male	80	1
58...	MIVIC HIROSE	PT196	Male	80	1

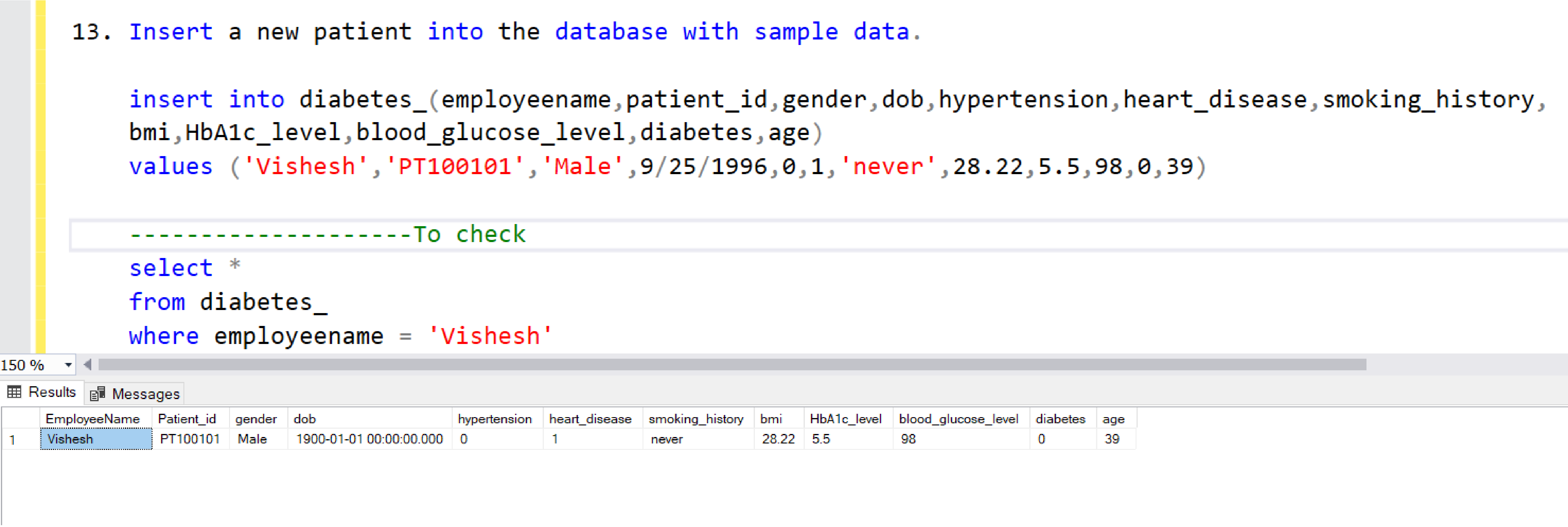


```
12.Update the smoking history of patients who are older than 50 to "Ex-smoker."

   update  diabetes_
   set smoking_history = 'Ex-smoker'
   where age > 33;

   ------To check "Ex-smoker"
   select smoking_history , count(*) as Total
    from diabetes_
    group by smoking_history
```

	smoking_history	Total
1	current	9282
2	Ex-smoker	50
3	not current	6446
4	former	9346
5	ever	4003
6	No Info	35800
7	never	35073

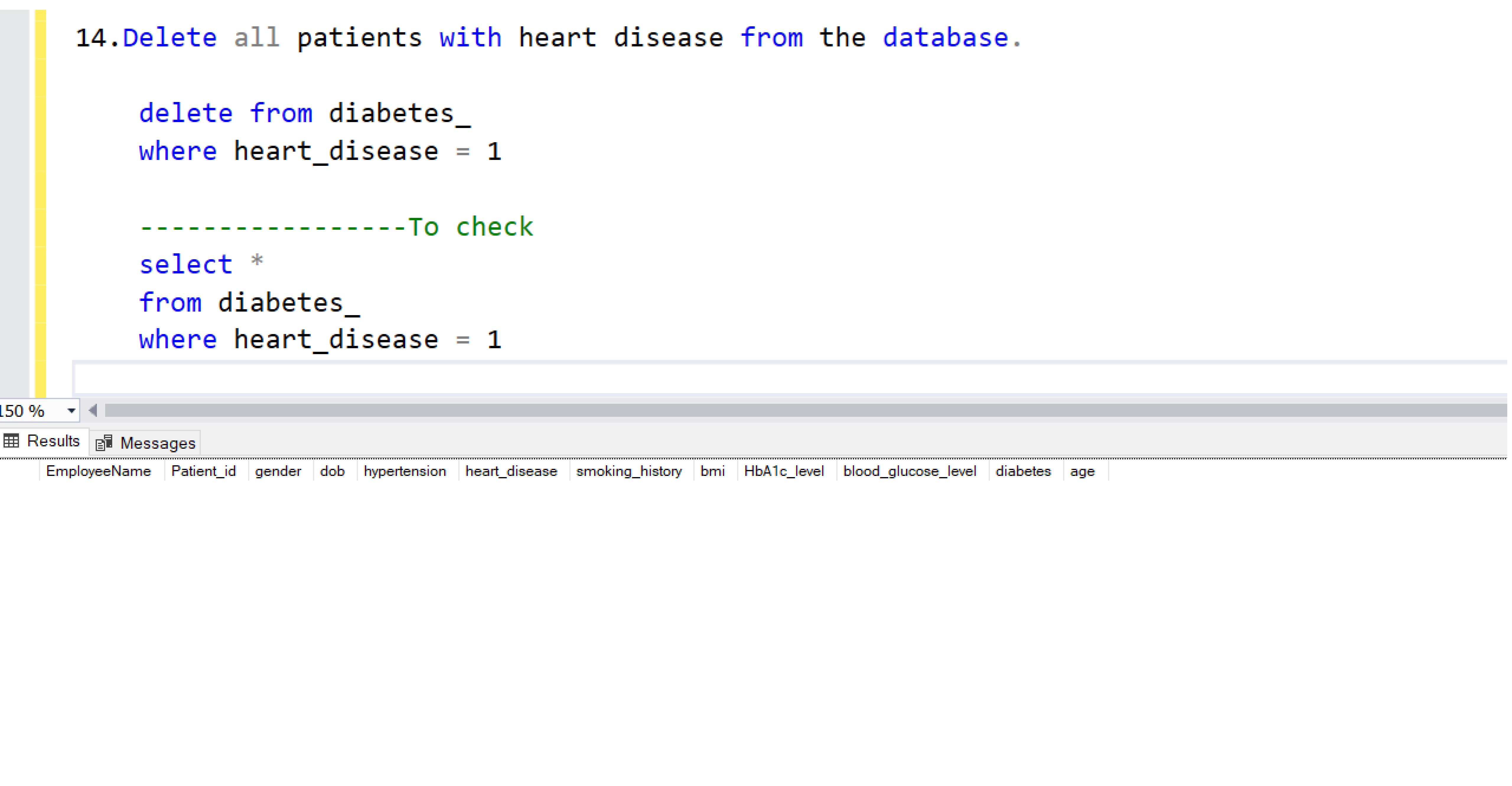


```
13. Insert a new patient into the database with sample data.

    insert into diabetes_(employeename,patient_id,gender,dob,hypertension,heart_disease,smoking_history,
    bmi,HbA1c_level,blood_glucose_level,diabetes,age)
    values ('Vishesh','PT100101','Male',9/25/1996,0,1,'never',28.22,5.5,98,0,39)

    --------------------To check
    select *
    from diabetes_
    where employeename = 'Vishesh'
```

	EmployeeName	Patient_id	gender	dob	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes	age
1	Vishesh	PT100101	Male	1900-01-01 00:00:00.000	0	1	never	28.22	5.5	98	0	39

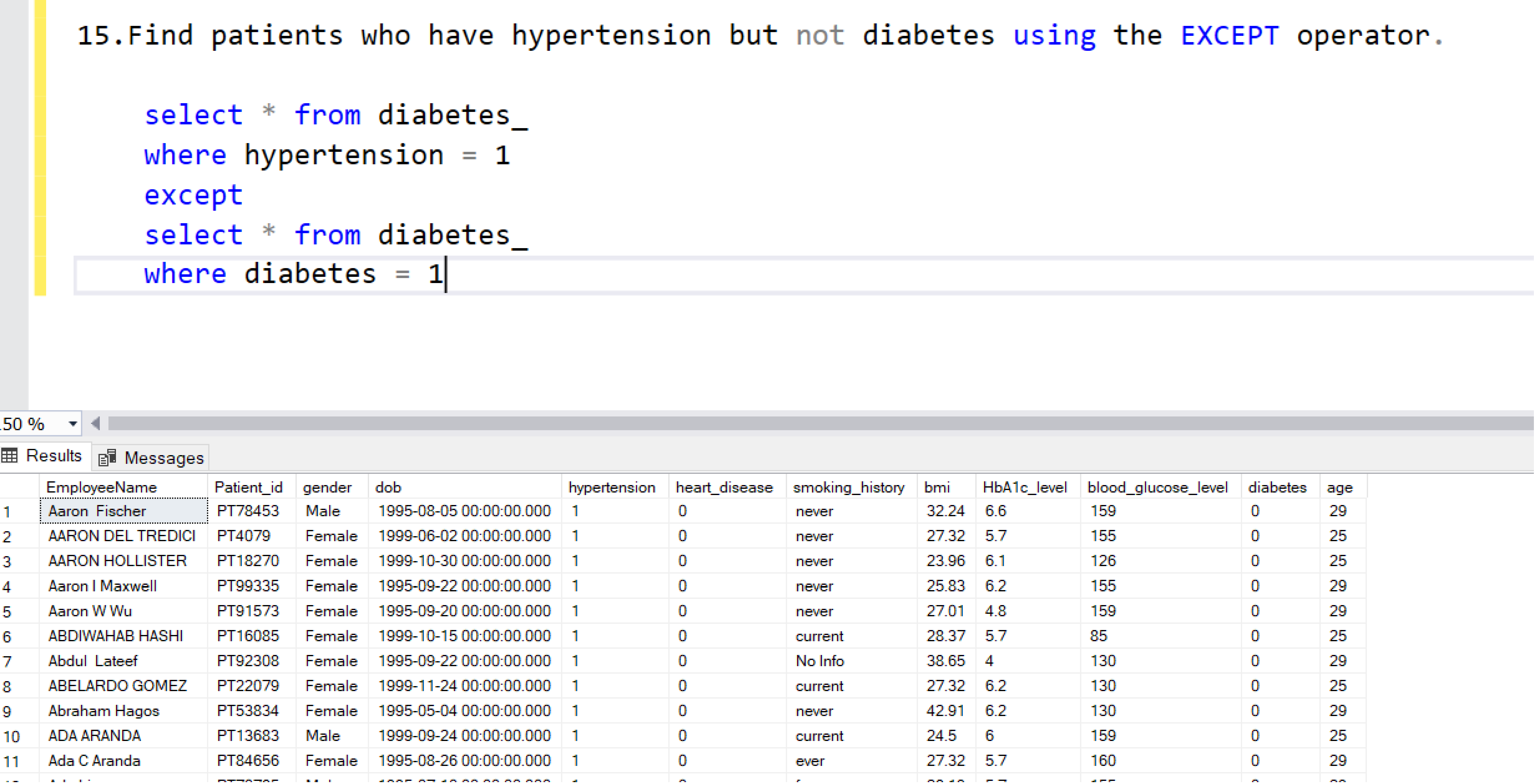


```
14.Delete all patients with heart disease from the database.

    delete from diabetes_
    where heart_disease = 1

    -----------------To check
    select *
    from diabetes_
    where heart_disease = 1
```

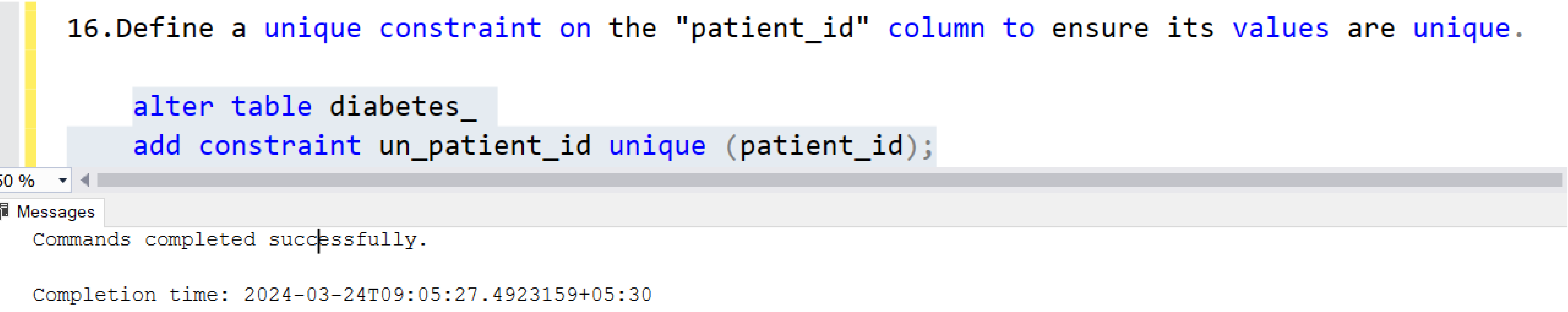
EmployeeName	Patient_id	gender	dob	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes	age



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	dob	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes	age
1	Aaron Fischer	PT78453	Male	1995-08-05 00:00:00.000	1	0	never	32.24	6.6	159	0	29
2	AARON DEL TREDICI	PT4079	Female	1999-06-02 00:00:00.000	1	0	never	27.32	5.7	155	0	25
3	AARON HOLLISTER	PT18270	Female	1999-10-30 00:00:00.000	1	0	never	23.96	6.1	126	0	25
4	Aaron I Maxwell	PT99335	Female	1995-09-22 00:00:00.000	1	0	never	25.83	6.2	155	0	29
5	Aaron W Wu	PT91573	Female	1995-09-20 00:00:00.000	1	0	never	27.01	4.8	159	0	29
6	ABDIWAHAB HASHI	PT16085	Female	1999-10-15 00:00:00.000	1	0	current	28.37	5.7	85	0	25
7	Abdul Lateef	PT92308	Female	1995-09-22 00:00:00.000	1	0	No Info	38.65	4	130	0	29
8	ABELARDO GOMEZ	PT22079	Female	1999-11-24 00:00:00.000	1	0	current	27.32	6.2	130	0	25
9	Abraham Hagos	PT53834	Female	1995-05-04 00:00:00.000	1	0	never	42.91	6.2	130	0	29
10	ADA ARANDA	PT13683	Male	1999-09-24 00:00:00.000	1	0	current	24.5	6	159	0	25
11	Ada C Aranda	PT84656	Female	1995-08-26 00:00:00.000	1	0	ever	27.32	5.7	160	0	29

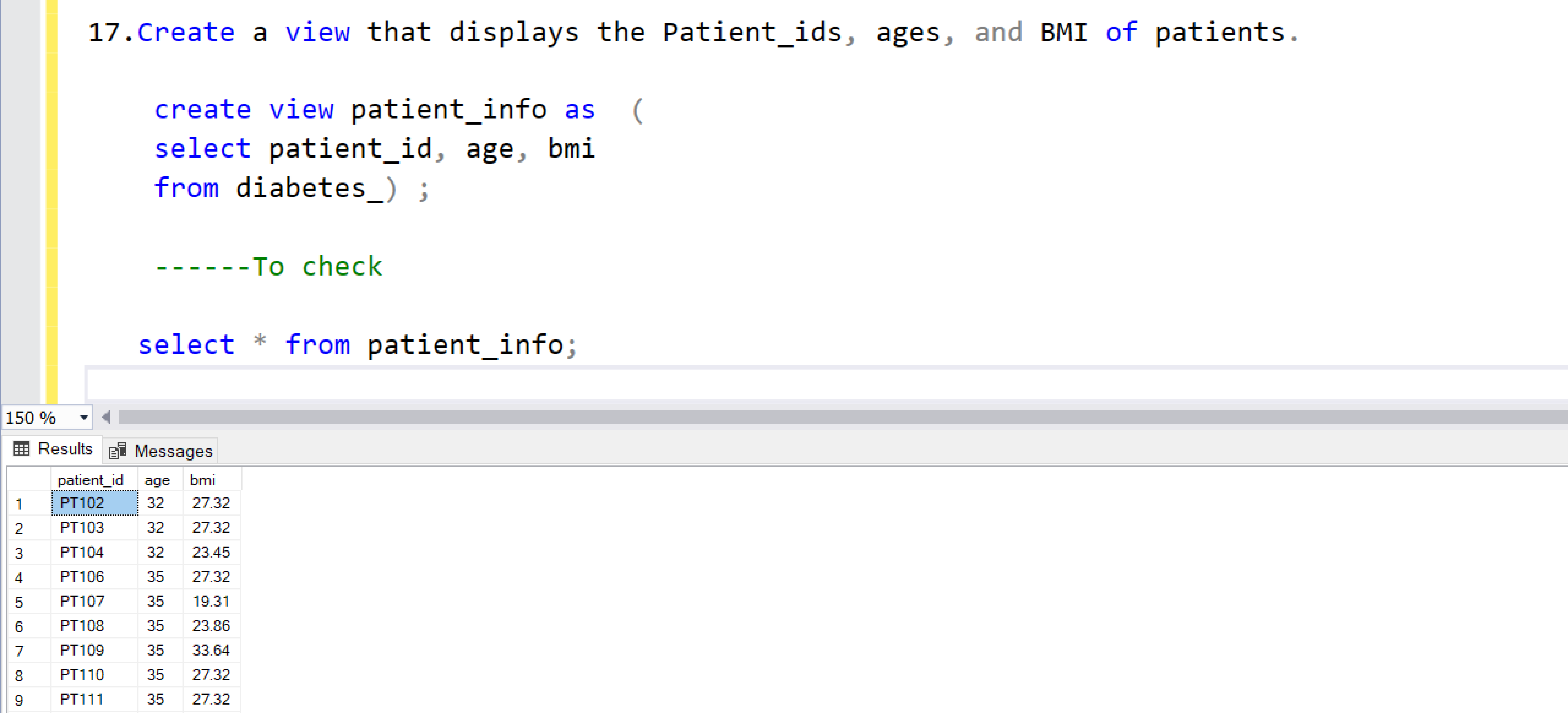


```
16.Define a unique constraint on the "patient_id" column to ensure its values are unique.

    alter table diabetes_
    add constraint un_patient_id unique (patient_id);
```

```
Commands completed successfully.

Completion time: 2024-03-24T09:05:27.4923159+05:30
```



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

    create view patient_info as  (
    select patient_id, age, bmi
    from diabetes_) ;

    ------To check

   select * from patient_info;
```

	patient_id	age	bmi
1	PT102	32	27.32
2	PT103	32	27.32
3	PT104	32	23.45
4	PT106	35	27.32
5	PT107	35	19.31
6	PT108	35	23.86
7	PT109	35	33.64
8	PT110	35	27.32
9	PT111	35	27.32

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

Redundancy means having multiple copies of the same data in the database. This problem arises when a database is not normalized.

1. Normalization : Normalization is a database design technique that involves efficiently organizing data to eliminate data redundancy and correct data dependency.
2. Use of Primary key : Ensure each table has primary key to uniquely identify each record. This will help to avoiding duplicate entries.
3. Foreign keys: Use foreign keys to establish relationship between tables. This maintains referential integrity and prevents inconsistencies.
4. Data types and Constraints: Choose appropriate data types for column to minimize storage space.
5. Composite keys: In case where a combination of columns can uniquely identify a record, consider using a composite key instead of a single column as the primary key.

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

1. Use Indexes –
   Identify columns frequently used in WHERE clause ,JOIN conditions & ORDER BY clauses.
2. Optimize Joins –
   Use INNER JOINs instead of OUTER JOINs when possible.
3. Reduce Data Retrieval-
   Retrieve only the necessary columns in SELECT statements rather than using SELECT *.
4. Filter Data Efficiently
   Use WHERE clauses to filter data early in the query execution process, reducing the amount of data processed.
5. Optimize Aggregation and Grouping
   Use appropriate aggregate functions e.g.- SUM,AVG.