

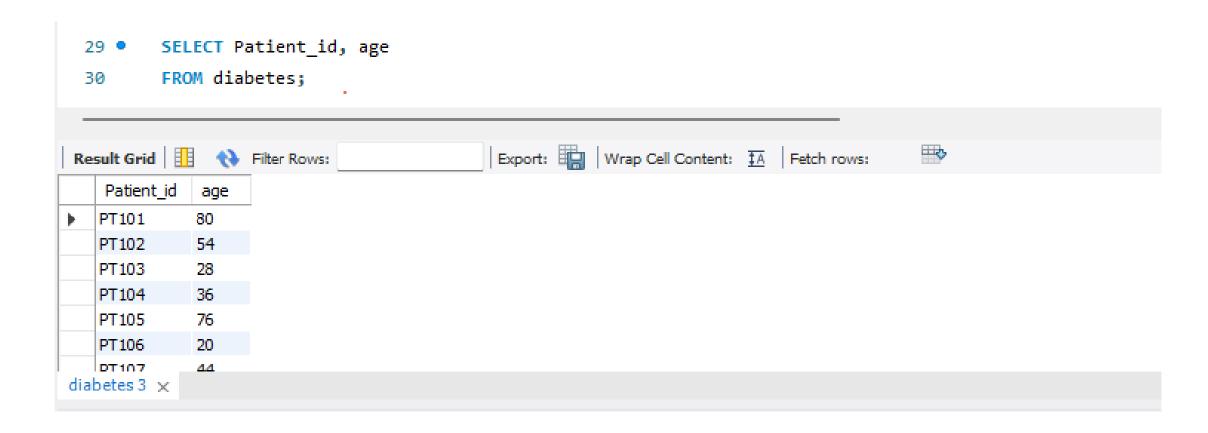
DATA ANALYST INTERNSHIP

DIABETES PREDICTION ANALYSIS

BY ANKUSH VERMA in



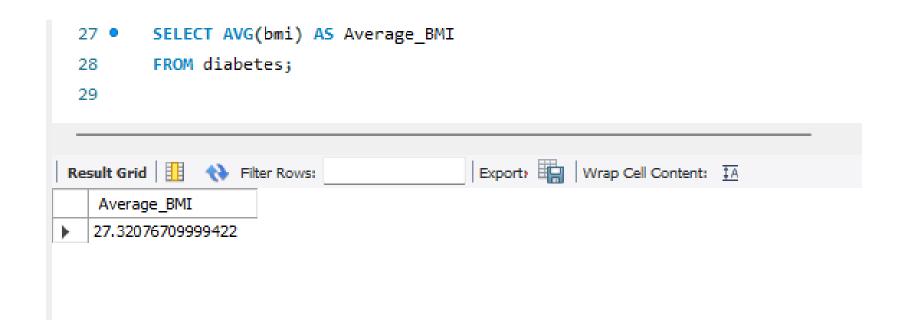
QI. RETRIEVE THE PATIENT_ID AND AGES OF ALL PATIENTS



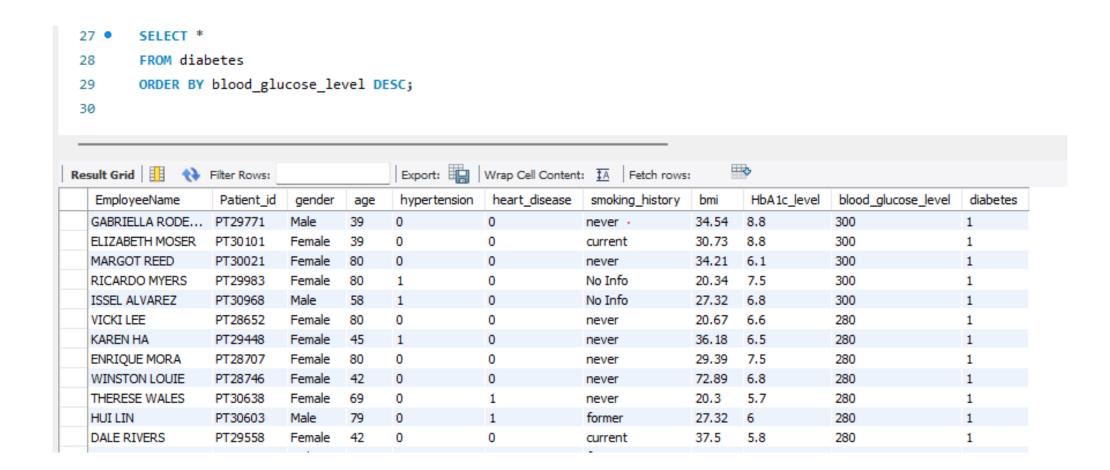
Q2. SELECT ALL FEMALE PATIENTS WHO ARE OLDER THAN 40

```
SELECT *
  29 •
  30
          FROM diabetes
          WHERE gender = "Female" AND age > 40;
 31
                                                                                                   4
Result Grid
                Filter Rows:
                                                             Wrap Cell Content: $\overline{A}$
                                                                                    Fetch rows:
                                                                                                          HbA1c_level
                                                                                                                       blood glucose level
                                                                                                                                           diabetes
    EmployeeName
                          Patient_id
                                      gender
                                                      hypertension
                                                                   heart disease
                                                                                  smoking history
                                                                                                  bmi
                                               age
   NATHANIEL FORD
                          PT101
                                     Female
                                                                                                  25, 19
                                                                                                          6.6
                                                                                                                       140
                                                                                                                                           0
                                                                                  never
                                              54
   GARY JIMENEZ
                          PT102
                                                                                  No Info
                                                                                                         6.6
                                     Female
                                                                                                  27.32
                                                                                                                                           0
   ALSON LEE
                          PT107
                                     Female
                                                                                                  19.31
                                                                                                         6.5
                                                                                                                       200
                                                                                  never
   DAVID KUSHNER
                          PT108
                                     Female
                                              79
                                                                                 No Info
                                                                                                  23.86
                                                                                                         5.7
                                                                                                                       85
   ARTHUR KENNEY
                          PT111
                                     Female
                                                                                                  27.32
                                                                                                                       85
                                                                                  never
   PATRICIA JACKSON
                          PT112
                                                                                                  54.7
                                                                                                                                           0
                                     Female
                                                                                  former
   EDWADD HADDINGTON
                          DT113
                                     Famala
                                                                   n
                                                                                  former
                                                                                                  36.05 5
                                                                                                                       130
                                                                                                                                           Λ.
diabetes 4 x
```

Q3. CALCULATE THE AVERAGE BMI OF PATIENTS



Q4. LIST PATIENTS IN DESCENDING ORDER OF BLOOD GLUCOSE LEVELS



Q5. FIND PATIENTS WHO HAVE HYPERTENSION AND DIABETES

```
27 • SELECT *
28 FROM diabetes
29 WHERE hypertension=1 AND diabetes=1;
30
```

EmployeeName	Patient id	gender	age	hypertension	heart disease	smoking history	bmi	HbA1c level	blood_glucose_level	diabetes
	_		50	1	0			5.7		1
ONES WONG	PT139	Male		1	_	current	27.32		260	_
ATRIC STEELE	PT205	Female	80	1	0	never .	27.32	6.8	280	1
RTHUR STELLINI	PT343	Male	57	1	1	not current	27.77	6.6	160	1
HAD LAW	PT355	Male	63	1	0	ever	35.06	5.8	200	1
ATHERINE JAMES	PT451	Female	52	1	0	never	50.3	6.6	155	1
OHN HART	PT565	Male	48	1	0	current	36.12	6.8	140	1
OHN BARKER	PT567	Female	79	1	0	former	27.32	6.5	159	1
OBERT BONNET	PT632	Female	49	1	0	not current	36.93	8.8	155	1
ITANI BENJAMIN	PT727	Male	43	1	0	not current	40.86	6.6	159	1
ANNIE ADELMAN	PT828	Female	38	1	0	not current	27.32	6.1	160	1
OEL DELIZONNA	PT852	Female	28	1	0	never	20.09	6.6	200	1
AREN KUBICK	PT861	Male	59	1	0	ever	25.94	9	140	1
OEL DELIZONNA	PT852	Female	28	1 1 1	0	never	20.09	6.6	200	

Q6. DETERMINE THE NUMBER OF PATIENTS WITH HEART DISEASE

```
SELECT COUNT(*) AS Heart_Diseases_Patients
       FROM diabetes
 28
       WHERE heart disease=1;
 29
 30
 31
Export: Wrap Cell Content: TA
  Heart_Diseases_Patients
 3942
```

Q7. GROUP PATIENTS BY SMOKING HISTORY AND COUNT HOW MANY SMOKERS AND NON-SMOKERS THERE ARE

```
SELECT smoking history, COUNT(*) AS COUNT
 28
         FROM diabetes
         WHERE smoking history="current" OR smoking history="never"
 29
 30
         GROUP BY smoking history;
 31
Result Grid
               ♦ Filter Rows:
                                                         Wrap Cell Content: $\frac{1}{4}$
   smoking_history
                  COUNT
                  35095
  never
                  9286
  current
```

Q8. RETRIEVE THE PATIENT_IDS OF PATIENTS WHO HAVE A BMI GREATER THAN THE AVERAGE BMI

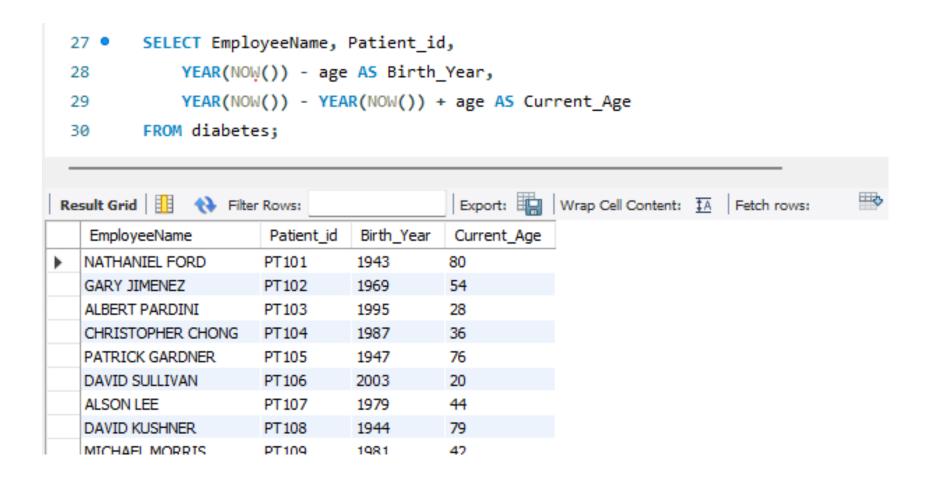
```
# AVG BMI is 27.32076709999422
         SELECT Patient id, bmi
         FROM diabetes

⊖ WHERE bmi > (
 31
              SELECT avg(bmi)
              FROM diabetes
32
 33
              );
 34
Result Grid
                                            Export: Wrap Cell Content: TA F
              Filter Rows:
   Patient_id
             bmi
  PT109
            33.64
  PT112
            54.7
  PT113
            36.05
            30.36
  PT117
  PT121
            36.38
  PT124
            27.94
  PT126
             33.76
```

Q9. FIND THE PATIENT WITH THE HIGHEST HBAIC LEVEL AND THE PATIENT WITH THE LOWEST HBAICLEVEL

```
SELECT EmployeeName, Patient id, HbA1c level as MAX HbA1c level
                                                                                             SELECT EmployeeName, Patient id, HbA1c level as Min HbA1c level
         FROM diabetes
 28
                                                                                             FROM diabetes
         ORDER BY HbA1c level DESC
                                                                                             ORDER BY HbA1c level ASC
         LIMIT 1;
                                                                                             LIMIT 1;
                                                                                                                               Export: Wrap Cell Content: A Fetch rows:
                                                                                    Result Grid
                                                                                                   ♦ Filter Rows:
Result Grid Filter Rows:
                                            Export: Wrap Cell Content: $\overline{A}$ Fetch row
                                                                                                      Patient id Min HbA1c level
                               MAX_HbA1c_level
                      Patient_id
   EmployeeName
                                                                                      ELLEN MOFFATT
                                                                                                     PT120
                                                                                                                3.5
  MICHAEL THOMPSON
                     PT141
```

Q10. CALCULATE THE AGE OF PATIENTS IN YEARS (ASSUMING THE CURRENT DATE AS OF NOW)



QII. 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) AS blood_glucose_level_rank_as_per_gender
FROM diabetes;

Re	Result Grid									
'	Patient_id	gender	blood_glucose_level	blood_glucose_level_rank_as_per_gender						
	PT47949	Female	80	1						
	PT47157	Female	80	1						
	PT47163	Female	80	1						
	PT47171	Female	80	1						
	PT49143	Female	80	1						
	PT46271	Female	85	4199						
	PT46110	Female	85	4199						
	PT48066	Female	85	4199						

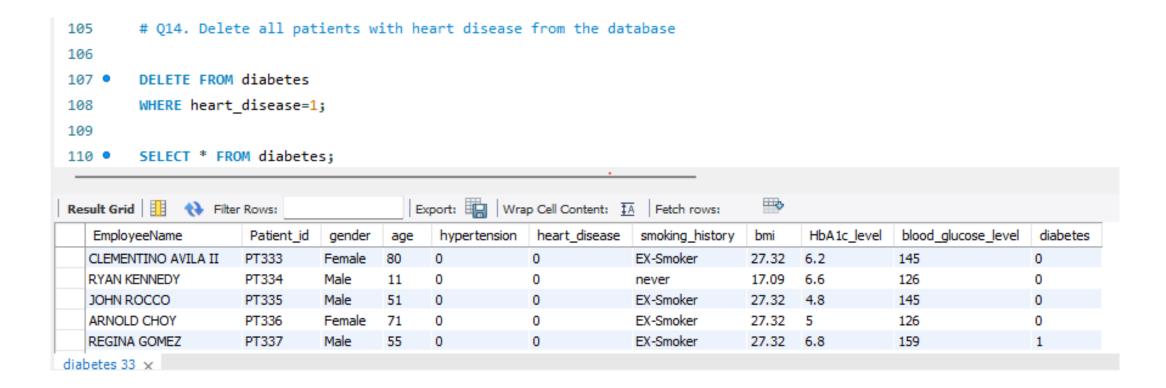
Q12. UPDATE THE SMOKING HISTORY OF PATIENTS WHO ARE OLDER THAN 50 TO "EX-SMOKER"

```
SET SQL SAFE UPDATES = 0;
         UPDATE diabetes
 28 •
         SET smoking history = "EX-Smoker"
 29
         WHERE age>50;
 30
 31
         SELECT * FROM diabetes;
Export: Wrap Cell Content: $\frac{1}{4}$
                                                                              Fetch rows:
   EmployeeName
                        Patient id
                                  gender
                                                              heart disease
                                                                             smoking history
                                                                                            bmi
                                                                                                    HbA1c level
                                                                                                                blood glucose level
                                                                                                                                   diabetes
                                                  hypertension
  NATHANIEL FORD
                       PT101
                                  Female
                                                                            EX-Smoker
                                                                                            25.19
                                                                                                                140
                                                                                            27.32
                       PT102
                                          54
                                                                            EX-Smoker
                                                                                                   6.6
  GARY JIMENEZ
                                  Female
                                                                                                                80
                                                                                                                                  0
  ALBERT PARDINI
                       PT103
                                  Male
                                           28
                                                                                            27.32
                                                                                                                158
                                                                            never
  CHRISTOPHER CHONG
                       PT104
                                  Female
                                          36
                                                                            current
                                                                                            23.45
                                                                                                                155
  PATRICK GARDNER
                       PT105
                                  Male
                                           76
                                                                                                                155
                                                                            EX-Smoker
                                                                                            20.14
                                                                                            27.32
  DAVID SULLIVAN
                       PT106
                                          20
                                                                                                  6.6
                                                                                                                85
                                  Female
                                                                            never
  ALSON LEE
                       PT107
                                  Female
                                                                                            19.31
                                                                                                   6.5
                                                                                                                200
                                                                            never
  DAVID KUSHNER
                       PT108
                                  Female
                                          79
                                                                            EX-Smoker
                                                                                            23,86
                                                                                                   5.7
                                                                                                                85
                       PT109
  MICHAEL MORRIS
                                  Male
                                                                            never
                                                                                            33.64
                                                                                                   4.8
                                                                                                                145
  JOANNE HAYES-WHITE PT110
                                  Female
                                          32
                                                                                            27.32
                                                                                                  - 5
                                                                                                                100
                                                                            never
```

Q13. INSERT A NEW PATIENT INTO THE DATABASE WITH SAMPLE DATA

```
34 •
         INSERT INTO diabetes
 35
         VALUES ("DAVID WARNER", "PT100101", "Male", 35, 0, 0, "No Info", 33.01, 5.1, 100, 0);
 36
         SELECT * FROM diabetes
 37 •
 38
         LIMIT 100102;
Export: Wrap Cell Content: $\frac{1}{4}
   EmployeeName
                       Patient id
                                 gender
                                                                           smoking history
                                                                                                 HbA1c level
                                                                                                             blood_glucose_level
                                                                                                                                diabetes
                                                hypertension
                                                            heart_disease
                                                                                          bmi
                                          age
  Antoinette L Wells
                                                                          No Info
                       PT100097
                                 Female
                                                                                          17.37
                                                                                                             100
  Richard D Swart
                       PT100098
                                 Male
                                                                          EX-Smoker
                                                                                         27.83
                                                                                                             155
  Vivian Chu
                       PT100099
                                 Female
                                         24
                                                                                         35.42 4
                                                                                                             100
                                                                          never
  Savitree Satram
                       PT100100
                                 Female
                                                                          EX-Smoker
                                                                                          22.43
                                                                                                             90
                                 Male
                                                                          No Info
                                                                                                5.1
  DAVID WARNER
                       PT100101
                                         35
                                                                                          33.01
                                                                                                             100
                                                                                                                               0
```

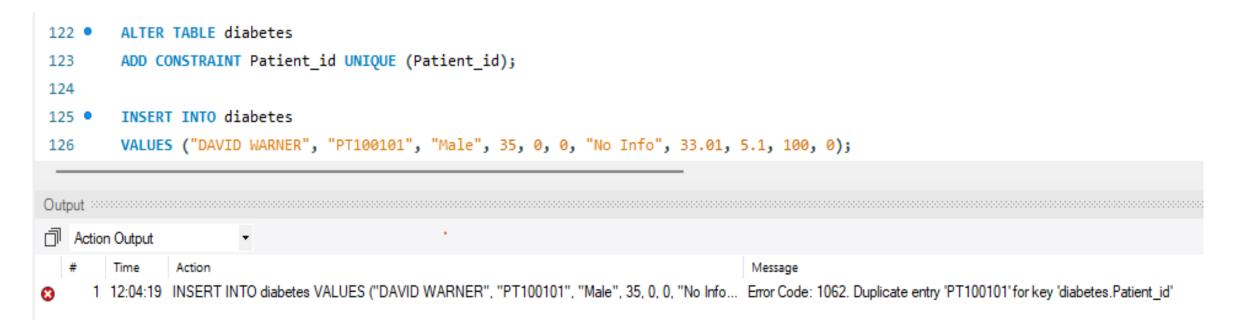
Q14. DELETE ALL PATIENTS WITH HEART DISEASE FROM THE DATABASE



Q15. FIND PATIENTS WHO HAVE HYPERTENSION BUT NOT DIABETES USING THE EXCEPT OPERATOR

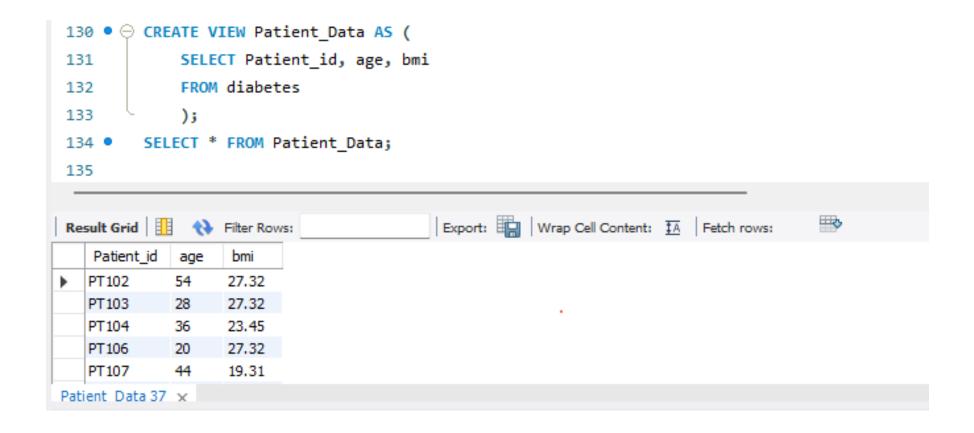
```
112
         # Q15. Find patients who have hypertension but not diabetes using the EXCEPT operator
113 •
         SELECT Patient id, hypertension, diabetes
114
         FROM diabetes
         WHERE hypertension=1
115
116 🖾
       EXCEPT
         SELECT Patient id, hypertension, diabetes
117
        FROM diabetes WHERE diabetes=1;
118
119
Result Grid Filter Rows:
                                            Export: Wrap Cell Content: $\frac{1}{4}$
   Patient id hypertension
                         diabetes
   PT958
                         0
   PT960
   PT972
   PT976
   PT980
                         0
Described as a
```

Q16. DEFINE A UNIQUE CONSTRAINT ON THE "PATIENT_ID" COLUMN TO ENSURE ITS VALUES ARE UNIQUE



Adding a UNIQUE constraint to 'Patient_id' means that attempting to insert a duplicate 'Patient_id' will result in an error.

Q17. CREATE A VIEW THAT DISPLAYS THE PATIENT_IDS, AGES, AND BMI OF PATIENTS



Q18. SUGGEST IMPROVEMENTS IN THE DATABASE SCHEMA TO REDUCE DATA REDUNDANCY AND IMPROVE DATA INTEGRITY

To Reduce Data Redundancy:

- I) Normalization: Split 'diabetes' table into 'Patients' and 'HealthRecords' tables, linking them with a foreign key on Patient_id.
- 2) Use Enumerations: Replace VARCHAR with ENUM for categorical data (e.g., gender, smoking_history) to enhance consistency.
- 3) Avoid Repeating Groups: Create a separate table for health records to establish a one-to-many relationship, preventing repeated health-related information in the main 'Patients' table.

To Improve Data Integrity:

- I) Implement Constraints: Enforce NOT NULL, UNIQUE, and FOREIGN KEY constraints for essential fields and to prevent duplicates.
- 2) Data Validation with Check Constraints: Use CHECK constraints for conditions like valid age range and blood_glucose_level values.
- **3) Triggers for Automation:** Implement triggers to automate actions, such as updating timestamps for health record modifications, aiding in data auditing.

Q19. EXPLAIN HOW YOU CAN OPTIMIZE THE PERFORMANCE OF SQL QUERIES ON THIS DATASET

- 1) Indexing: Create indexes on frequently queried columns like Patient_id, age, and diabetes for faster retrieval.
- 2) Use Proper Joins: Optimize JOIN operations, ensuring efficient linking between tables.
- 3) Limit SELECT Columns: Retrieve only necessary columns to minimize data transfer and improve query speed.
- 4) Update Statistics: Regularly update database statistics to help the query optimizer generate efficient execution plans.
- 5) Partitioning: Consider partitioning large tables based on certain criteria (e.g., date) to enhance query performance.