

Healthcare Analysis using SQL

Contents

Project Summary:.....	2
Objective:	2
Dataset:	3
Query Analysis and Insights:	4
DEMOGRAPHICS	4
DIAGNOSIS	5
APPOINTMENT TIME DISTRIBUTION	7
COMMON LAB TEST	8
LAB RESULT ANALYSIS.....	9
PATIENT RISK CATEGORY	10
PATIENT VISIT HISTORY.....	11
Conclusion.....	12

Project Summary:

This project focuses on analysing data to address the hospital's concerns regarding resource allocation and the effectiveness of patient care across various departments. The goal is to uncover actionable insights into patient demographics, diagnostic trends, and visit patterns to optimize service delivery.

Objective:

1) What is the demographic profile of the patient population, including age and gender distribution?

Pediatric: less than 18 years old

Adult: Between 18 to 64 years old

Senior: Over 65 years old

2) Which diagnoses are most prevalent among patients and how do they vary across the different demographic groups, including gender and age?

3) What are the most common appointment times throughout the day, and how does the distribution of appointment times vary across different hours?

4) What are the most ordered lab tests?

5) Typically, fasting blood sugar levels fall between 70-100 mg/dL. Identify patients whose lab results are outside this normal range to implement early intervention.

6) Assess how many patients are considered High, Medium, and Low Risk. High Risk: patients who are smokers and have been diagnosed with either hypertension or diabetes.

7) Find out information about patients who had multiple visits within 30 days of their previous medical.

Dataset:

Below are details of the datasets:

- Patients: patient id, patient name, date of birth, gender, address
- Outpatient Visits: visit id, patient id, visit date, name of doctor, reason for visit, diagnosis prescribed medication, smoking status
- Lab Results: result id, visit id, test name, test date, result value
- Hospital Records: patient id, patient name, BMI, family history, days in hospital
- Appointment: visit id, patient id, department name, patient name, appointment time, arrival time, appointment date, admission time

Query Analysis and Insights:

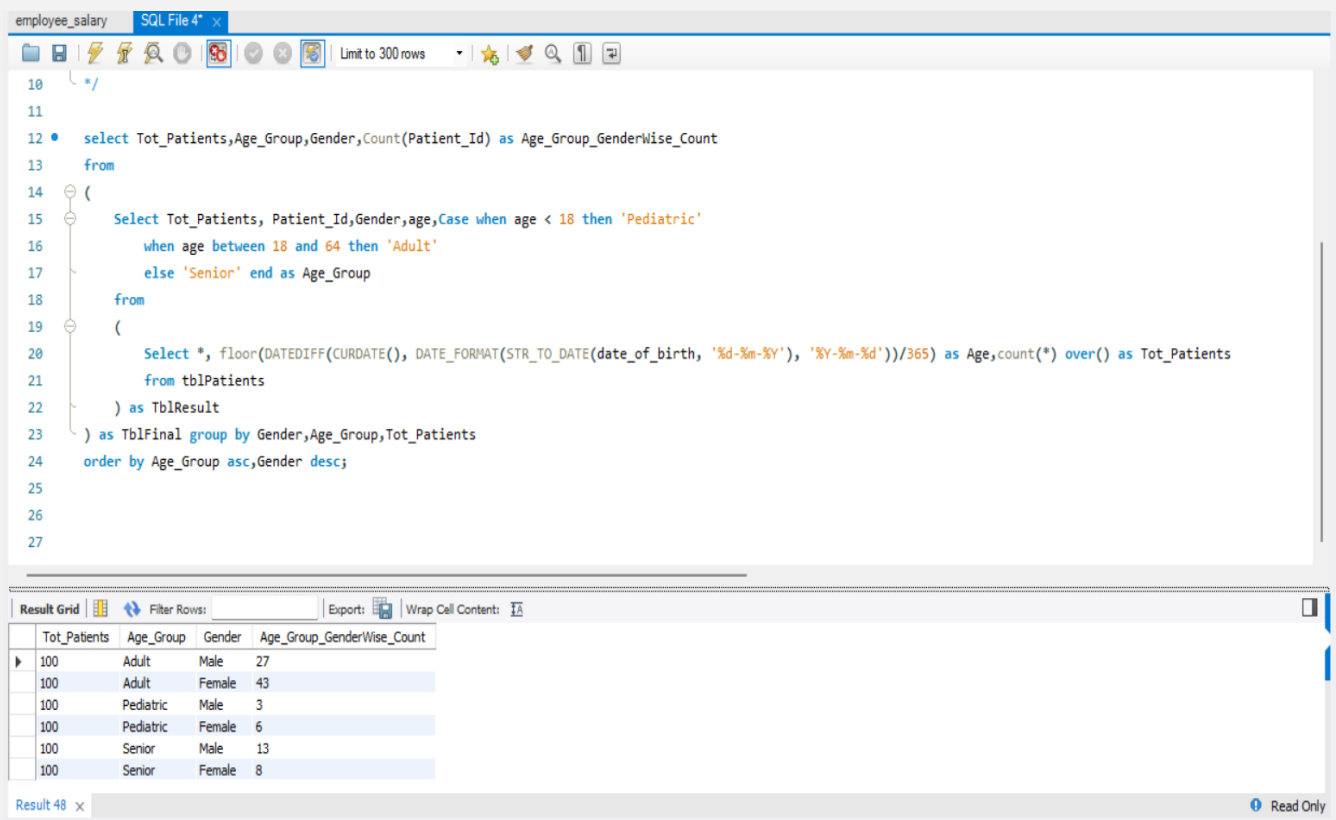
DEMOGRAPHICS

1) What is the demographic profile of the patient population, including age and gender distribution?

- Pediatric: less than 18 years old
- Adult: Between 18 to 64 years old
- Senior: Over 65 years old

I grouped the records by Age group and Gender and observed that Adult Females represent the largest segment in the patient dataset, followed by Adult Males.

QUERY & OUTPUT



The screenshot displays a SQL IDE window titled "employee_salary" with a tab for "SQL File 4". The query editor contains the following SQL code:

```
10  */
11
12  • select Tot_Patients, Age_Group, Gender, Count(Patient_Id) as Age_Group_GenderWise_Count
13  from
14  (
15      Select Tot_Patients, Patient_Id, Gender, age, Case when age < 18 then 'Pediatric'
16          when age between 18 and 64 then 'Adult'
17          else 'Senior' end as Age_Group
18  from
19  (
20      Select *, floor(DATEDIFF(CURDATE(), DATE_FORMAT(STR_TO_DATE(date_of_birth, '%d-%m-%Y'), '%Y-%m-%d'))/365) as Age, count(*) over() as Tot_Patients
21      from tblPatients
22  ) as TblResult
23  ) as TblFinal group by Gender, Age_Group, Tot_Patients
24  order by Age_Group asc, Gender desc;
25
26
27
```

Below the query editor, the "Result Grid" shows the output of the query. The grid has columns: Tot_Patients, Age_Group, Gender, and Age_Group_GenderWise_Count. The results are as follows:

Tot_Patients	Age_Group	Gender	Age_Group_GenderWise_Count
100	Adult	Male	27
100	Adult	Female	43
100	Pediatric	Male	3
100	Pediatric	Female	6
100	Senior	Male	13
100	Senior	Female	8

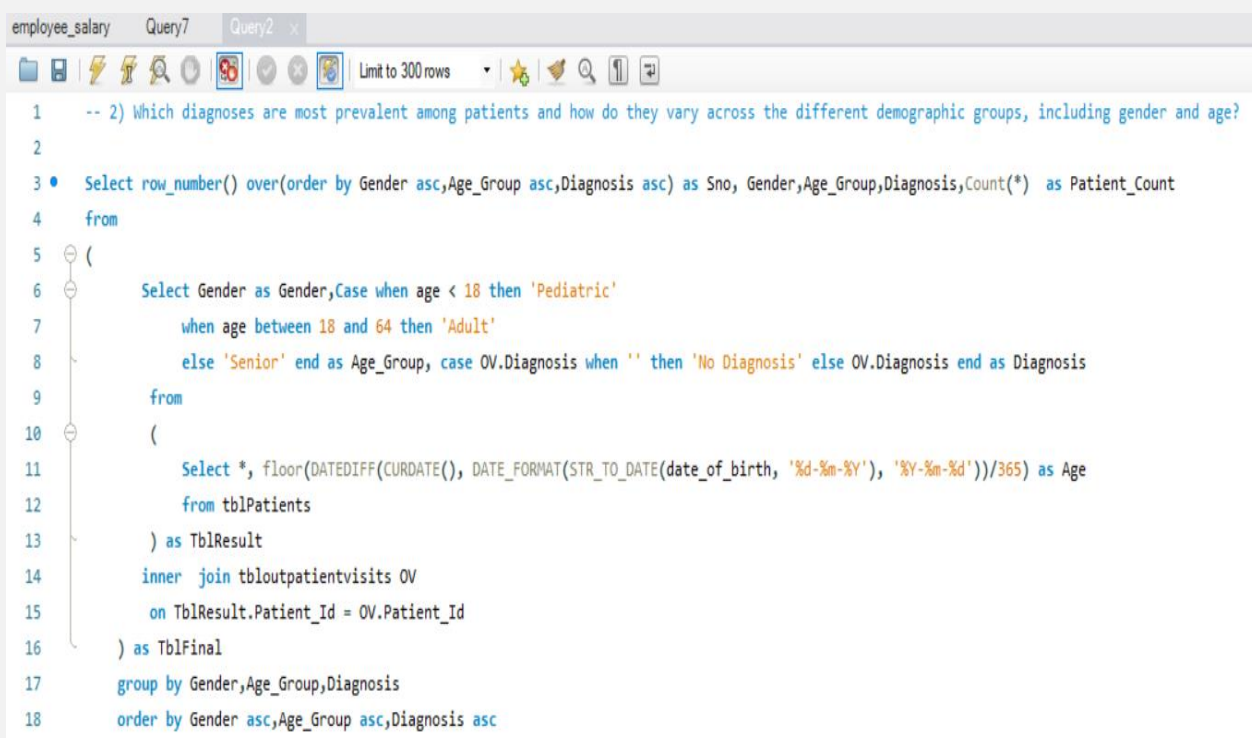
The status bar at the bottom indicates "Result 48" and "Read Only".

DIAGNOSIS

2) Which diagnoses are most prevalent among patients and how do they vary across the different demographic groups, including gender and age?

109 adult females and 25 senior females were left without a diagnosis, while 68 adult males and 51 senior males also had no recorded diagnosis. In addition, the most common diagnosis among adult males and females is diabetes, while respiratory illness is prevalent among seniors.

QUERY



```
1  -- 2) Which diagnoses are most prevalent among patients and how do they vary across the different demographic groups, including gender and age?
2
3  • Select row_number() over(order by Gender asc, Age_Group asc, Diagnosis asc) as Sno, Gender, Age_Group, Diagnosis, Count(*) as Patient_Count
4  from
5  (
6      Select Gender as Gender, Case when age < 18 then 'Pediatric'
7          when age between 18 and 64 then 'Adult'
8          else 'Senior' end as Age_Group, case OV.Diagnosis when '' then 'No Diagnosis' else OV.Diagnosis end as Diagnosis
9      from
10     (
11         Select *, floor(DATEDIFF(CURDATE(), DATE_FORMAT(STR_TO_DATE(date_of_birth, '%d-%m-%Y'), '%Y-%m-%d'))/365) as Age
12         from tblPatients
13     ) as TblResult
14     inner join tbloutpatientvisits OV
15     on TblResult.Patient_Id = OV.Patient_Id
16 ) as TblFinal
17 group by Gender, Age_Group, Diagnosis
18 order by Gender asc, Age_Group asc, Diagnosis asc
```

OUTPUT

FEMALE DISTRIBUTION

Result Grid					
Filter Rows:					
Export: Wrap Cell Content:					
Sno	Gender	Age_Group	Diagnosis	Patient_Count	
1	Female	Adult	Allergic Reaction	2	
2	Female	Adult	Common Cold	3	
3	Female	Adult	Diabetes	21	
4	Female	Adult	Food Poisoning	1	
5	Female	Adult	Hyperlipidemia	7	
6	Female	Adult	Hypertension	13	
7	Female	Adult	Hypothyroid	3	
8	Female	Adult	Migraine	5	
9	Female	Adult	Mild Concussion	1	
10	Female	Adult	Muscle Injury	12	
11	Female	Adult	No Diagnosis	113	
12	Female	Adult	Respiratory Illness	12	
13	Female	Adult	Respiratory infection	2	
14	Female	Pediatric	Bronchiolitis	3	
15	Female	Pediatric	Common Cold	2	
16	Female	Pediatric	Ear Infection	4	
17	Female	Pediatric	Gastroenteritis	1	
18	Female	Pediatric	No Diagnosis	14	
19	Female	Pediatric	Respiratory infection	2	
20	Female	Pediatric	Strep Throat	1	
21	Female	Senior	Common Cold	1	
22	Female	Senior	Diabetes	1	
23	Female	Senior	Mild Concussion	2	
24	Female	Senior	Muscle Injury	2	
25	Female	Senior	No Diagnosis	21	

MALE DISTRIBUTION

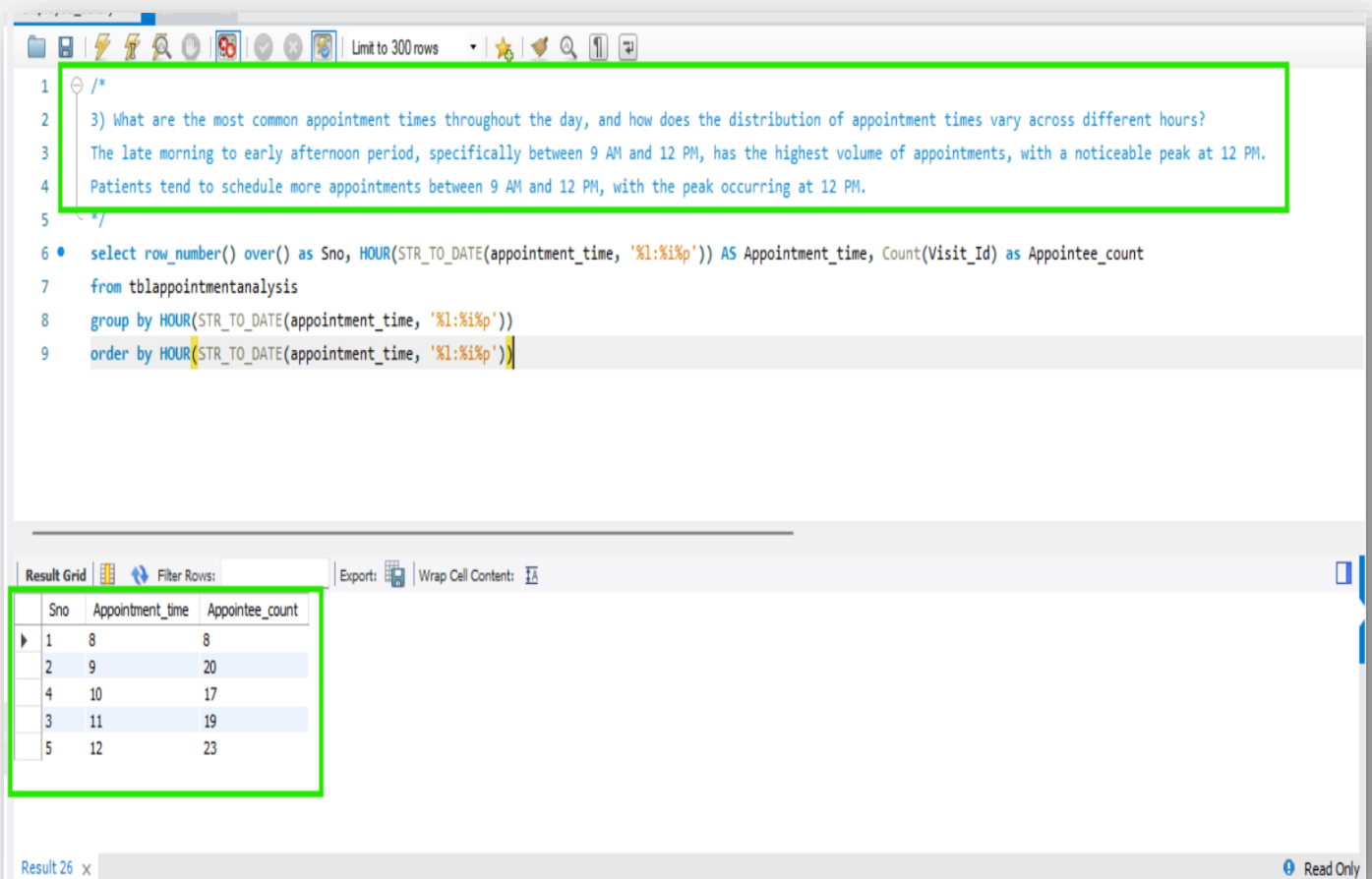
Result Grid					
Filter Rows:					
Export: Wrap Cell Content:					
Sno	Gender	Age_Group	Diagnosis	Patient_Count	
25	Female	Senior	No Diagnosis	21	
26	Female	Senior	Respiratory Illness	4	
27	Male	Adult	Allergic Reaction	5	
28	Male	Adult	Common Cold	4	
29	Male	Adult	Diabetes	11	
30	Male	Adult	Hyperlipidemia	2	
31	Male	Adult	Hypertension	9	
32	Male	Adult	Migraine	3	
33	Male	Adult	Muscle Injury	8	
34	Male	Adult	No Diagnosis	71	
35	Male	Adult	Respiratory Illness	6	
36	Male	Pediatric	Bronchiolitis	3	
37	Male	Pediatric	Common Cold	3	
38	Male	Pediatric	Ear Infection	2	
39	Male	Pediatric	Gastroenteritis	2	
40	Male	Pediatric	No Diagnosis	3	
41	Male	Pediatric	Strep Throat	2	
42	Male	Senior	Allergic Reaction	1	
43	Male	Senior	Diabetes	4	
44	Male	Senior	Hypertension	1	
45	Male	Senior	Migraine	1	
46	Male	Senior	Mild Concussion	1	
47	Male	Senior	Muscle Injury	2	
48	Male	Senior	No Diagnosis	48	

APPOINTMENT TIME DISTRIBUTION

3) What are the most common appointment times throughout the day, and how does the distribution of appointment times vary across different hours?

The late morning to early afternoon period, specifically between 9 AM and 12 PM, has the highest volume of appointments, with a noticeable peak at 12 PM. Patients tend to schedule more appointments between 9 AM and 12 PM, with the peak occurring at 12 PM.

QUERY & OUTPUT



The screenshot shows a SQL query editor with a query and its corresponding output grid. The query is a SQL statement that uses a window function to rank appointment times by count. The output grid displays the results of this query, showing the top 5 appointment times by count.

```
1 /*
2 3) What are the most common appointment times throughout the day, and how does the distribution of appointment times vary across different hours?
3 The late morning to early afternoon period, specifically between 9 AM and 12 PM, has the highest volume of appointments, with a noticeable peak at 12 PM.
4 Patients tend to schedule more appointments between 9 AM and 12 PM, with the peak occurring at 12 PM.
5 */
6 select row_number() over() as Sno, HOUR(STR_TO_DATE(appointment_time, '%l:%i%p')) AS Appointment_time, Count(Visit_Id) as Appointee_count
7 from tblappointmentanalysis
8 group by HOUR(STR_TO_DATE(appointment_time, '%l:%i%p'))
9 order by HOUR(STR_TO_DATE(appointment_time, '%l:%i%p'))
```

Sno	Appointment_time	Appointee_count
1	8	8
2	9	20
4	10	17
3	11	19
5	12	23

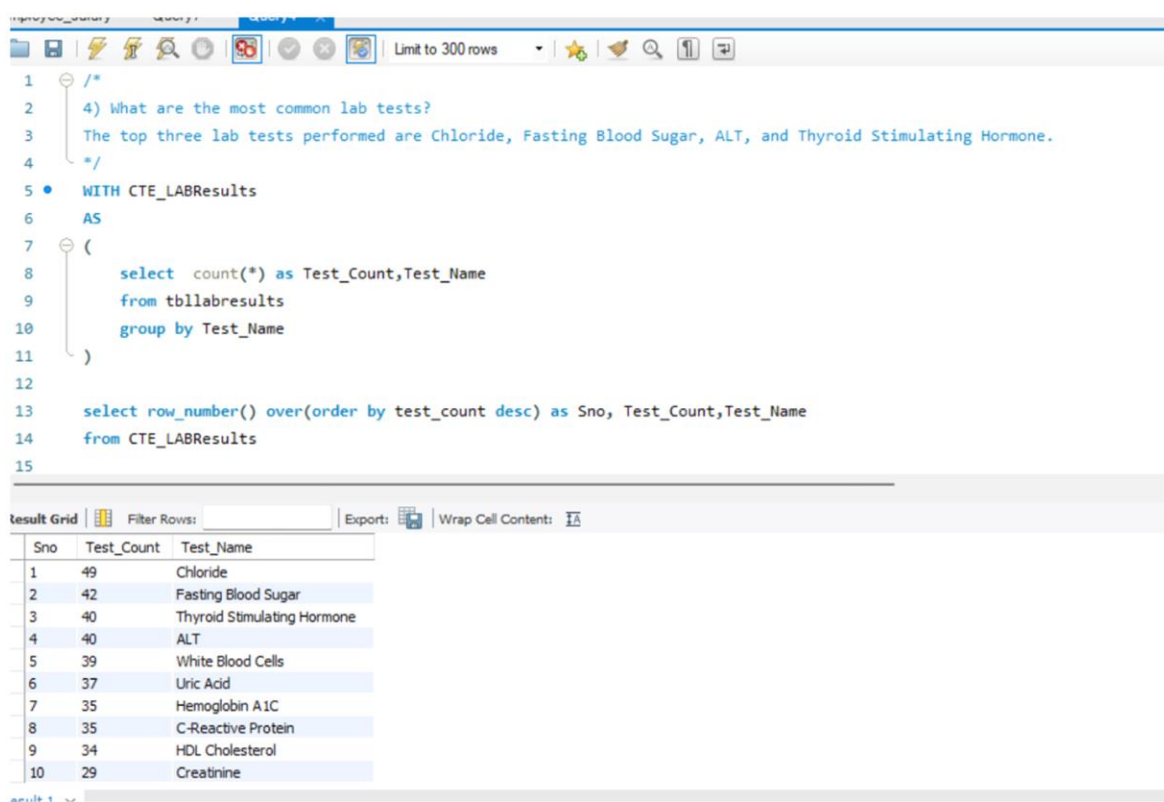
Result 26 x Read Only

COMMON LAB TEST

4) What are the most common lab tests?

The top three lab tests performed are Chloride, Fasting Blood Sugar, ALT, and Thyroid Stimulating Hormone.

QUERY & OUTPUT



```
1  /*
2  4) What are the most common lab tests?
3  The top three lab tests performed are Chloride, Fasting Blood Sugar, ALT, and Thyroid Stimulating Hormone.
4  */
5  WITH CTE_LABResults
6  AS
7  (
8      select count(*) as Test_Count,Test_Name
9      from tbllabresults
10     group by Test_Name
11 )
12
13 select row_number() over(order by test_count desc) as Sno, Test_Count,Test_Name
14 from CTE_LABResults
15
```

Sno	Test_Count	Test_Name
1	49	Chloride
2	42	Fasting Blood Sugar
3	40	Thyroid Stimulating Hormone
4	40	ALT
5	39	White Blood Cells
6	37	Uric Acid
7	35	Hemoglobin A1C
8	35	C-Reactive Protein
9	34	HDL Cholesterol
10	29	Creatinine

LAB RESULT ANALYSIS

5) Typically, fasting blood sugar levels fall between 70-100 mg/dL. Identify patients whose lab results are outside this normal range to implement early intervention.

These patients have fasting blood sugar levels outside the normal range and may benefit from further evaluation and early intervention to manage their blood sugar levels.

High Risk: patients who are smokers and have been diagnosed with either hypertension or diabetes.

Medium Risk: patients who are non-smokers and have been diagnosed with either hypertension or diabetes.

Low Risk: patients who do not fall into the High or Medium Risk categories. This includes patients who are not smokers and do not have a diagnosis of hypertension or diabetes.

The screenshot shows a SQL IDE window with a query editor and a results grid. The query editor contains a SQL query that filters for patients with fasting blood sugar levels outside the 70-100 mg/dL range. The results grid displays 10 rows of patient data, including their ID, name, and blood sugar result. A red box highlights the query text, and a green box highlights the results grid.

```
-- 5) Typically, fasting blood sugar levels fall between 70-100 mg/dL.
Identify patients whose lab results are outside this normal range to implement early intervention.
*/

use Capstoneprojects;
select * from tbllabresults;
select * from tblpatients;
select * from tblhospitalrecords;

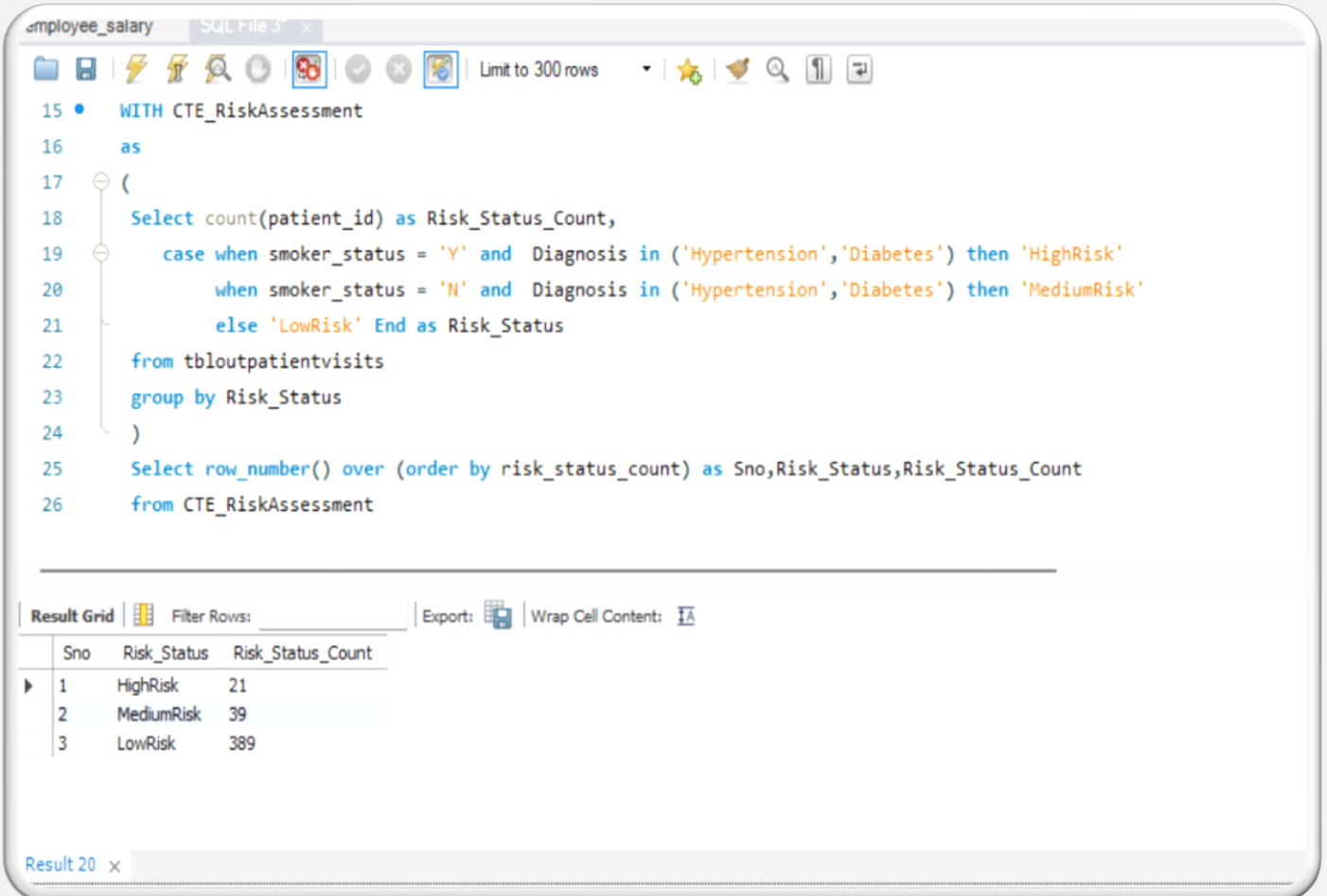
select row_number() over(order by p.patient_id) as Sno,p.patient_id,P.Patient_name,result_value
from tblpatients P
inner join tbloutpatientvisits OV on P.patient_Id = OV.patient_Id
inner join tbllabresults lr on Lr.Visit_id = OV.Visit_Id
where lr.test_name = 'Fasting Blood Sugar'
and result_value < 70 or result_value > 100
```

Sno	patient_id	Patient_name	result_value
1	521001	Emma Johnson	107
2	521001	Emma Johnson	101
3	521002	Michael Smith	103
4	521004	William Jones	105.82
5	521004	William Jones	110
6	521004	William Jones	101
7	521007	Isabella Taylor	102
8	521008	Bethany Clark	58.95
9	521009	Jose Gonzalez	120
10	521009	Jose Gonzalez	110

PATIENT RISK CATEGORY

6) Assess how many patients are considered High, Medium, and Low Risk.

There are 21 high-risk patients who are smokers and have been diagnosed with either Hypertension or Diabetes. Identifying these patients is crucial to provide them with appropriate treatment.



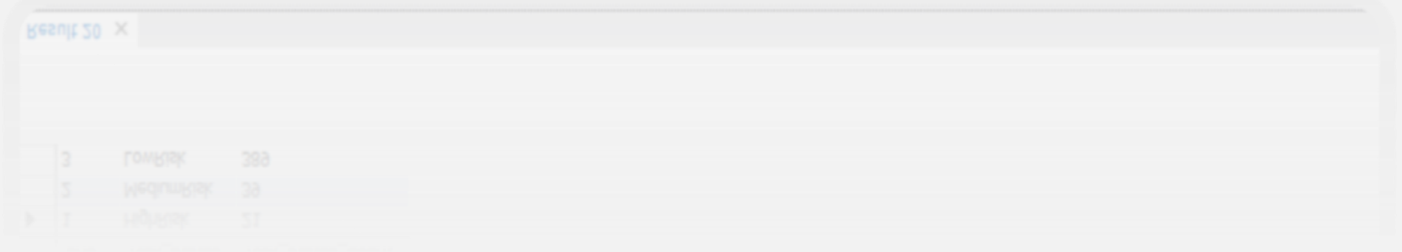
The screenshot displays a SQL IDE window titled "employee_salary" with a tab for "SQL File 3". The query editor contains the following SQL code:

```
15 • WITH CTE_RiskAssessment
16 as
17 (
18     Select count(patient_id) as Risk_Status_Count,
19     case when smoker_status = 'Y' and Diagnosis in ('Hypertension','Diabetes') then 'HighRisk'
20           when smoker_status = 'N' and Diagnosis in ('Hypertension','Diabetes') then 'MediumRisk'
21           else 'LowRisk' End as Risk_Status
22     from tbloutpatientvisits
23     group by Risk_Status
24 )
25 Select row_number() over (order by risk_status_count) as Sno,Risk_Status,Risk_Status_Count
26 from CTE_RiskAssessment
```

Below the query editor, the "Result Grid" shows the output of the query. The grid has columns for "Sno", "Risk_Status", and "Risk_Status_Count". The results are as follows:

Sno	Risk_Status	Risk_Status_Count
1	HighRisk	21
2	MediumRisk	39
3	LowRisk	389

The IDE interface includes a toolbar with various icons for file operations, a "Limit to 300 rows" dropdown, and a "Filter Rows" section. The "Result Grid" also features an "Export" button and a "Wrap Cell Content" checkbox.



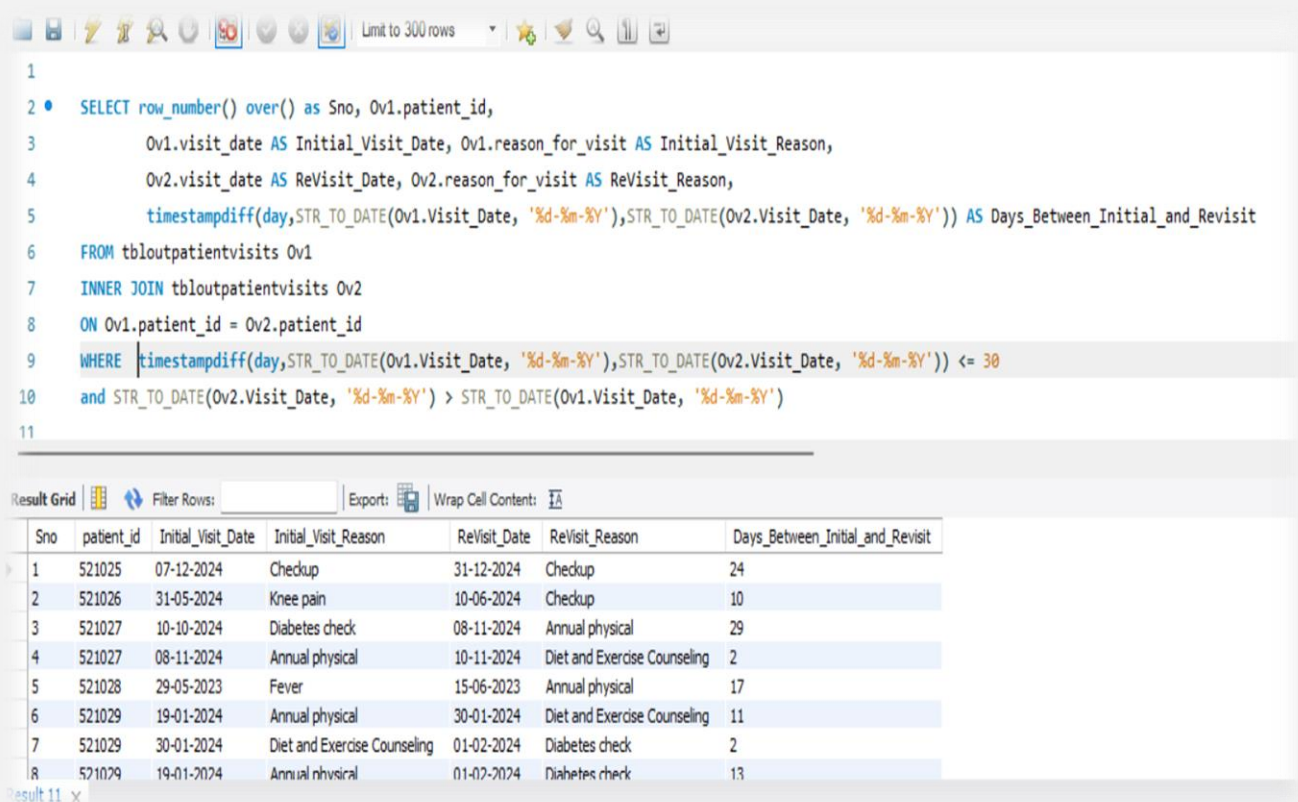
This block shows a faint, secondary instance of the SQL IDE window, mirroring the content of the first screenshot. It displays the same SQL query and the resulting data grid with three rows: HighRisk (21), MediumRisk (39), and LowRisk (389).

PATIENT VISIT HISTORY

7) Find out information about patients who had multiple visits within 30 days of their previous medical visit

- Identify those patients
- Date of initial visit
- Reason for the initial visit
- Readmission date
- Reason for readmission
- Number of days between the initial visit and readmission
- Readmission visit recorded must have happened after the initial visit

This analysis enables hospitals to pinpoint patients at risk for readmission, helping them enhance post-discharge care, minimize avoidable readmissions, and refine clinical protocols. Additionally, it highlights instances where the diagnosis from the initial visit may have had an adverse effect, leading to the need for readmission.



The screenshot shows a SQL query editor with a toolbar at the top. The query is as follows:

```
1
2 • SELECT row_number() over() as Sno, Ov1.patient_id,
3       Ov1.visit_date AS Initial_Visit_Date, Ov1.reason_for_visit AS Initial_Visit_Reason,
4       Ov2.visit_date AS ReVisit_Date, Ov2.reason_for_visit AS ReVisit_Reason,
5       timestampdiff(day,STR_TO_DATE(Ov1.Visit_Date, '%d-%m-%Y'),STR_TO_DATE(Ov2.Visit_Date, '%d-%m-%Y')) AS Days_Between_Initial_and_Revisit
6 FROM tbloutpatientvisits Ov1
7 INNER JOIN tbloutpatientvisits Ov2
8 ON Ov1.patient_id = Ov2.patient_id
9 WHERE timestampdiff(day,STR_TO_DATE(Ov1.Visit_Date, '%d-%m-%Y'),STR_TO_DATE(Ov2.Visit_Date, '%d-%m-%Y')) <= 30
10 and STR_TO_DATE(Ov2.Visit_Date, '%d-%m-%Y') > STR_TO_DATE(Ov1.Visit_Date, '%d-%m-%Y')
11
```

Below the query editor is the 'Result Grid' showing the results of the query. The grid has columns: Sno, patient_id, Initial_Visit_Date, Initial_Visit_Reason, ReVisit_Date, ReVisit_Reason, and Days_Between_Initial_and_Revisit. The results are as follows:

Sno	patient_id	Initial_Visit_Date	Initial_Visit_Reason	ReVisit_Date	ReVisit_Reason	Days_Between_Initial_and_Revisit
1	521025	07-12-2024	Checkup	31-12-2024	Checkup	24
2	521026	31-05-2024	Knee pain	10-06-2024	Checkup	10
3	521027	10-10-2024	Diabetes check	08-11-2024	Annual physical	29
4	521027	08-11-2024	Annual physical	10-11-2024	Diet and Exercise Counseling	2
5	521028	29-05-2023	Fever	15-06-2023	Annual physical	17
6	521029	19-01-2024	Annual physical	30-01-2024	Diet and Exercise Counseling	11
7	521029	30-01-2024	Diet and Exercise Counseling	01-02-2024	Diabetes check	2
8	521029	19-01-2024	Annual physical	01-02-2024	Diabetes check	13

Conclusion:

This analysis provides a comprehensive understanding of the patient population by demographic factors such as age and gender, offering insights into the prevalence of various diagnoses across different groups. It also identifies common appointment times, helping optimize scheduling, and highlights the most ordered lab tests, aiding in resource allocation. Additionally, by assessing fasting blood sugar levels, the hospital can take early intervention measures for patients outside the normal range. The risk categorization (High, Medium, and Low) allows for targeted healthcare strategies, while tracking patients with multiple visits within 30 days can help improve post-discharge care and reduce readmission rates.

By:

Manikantan
Database - MySql