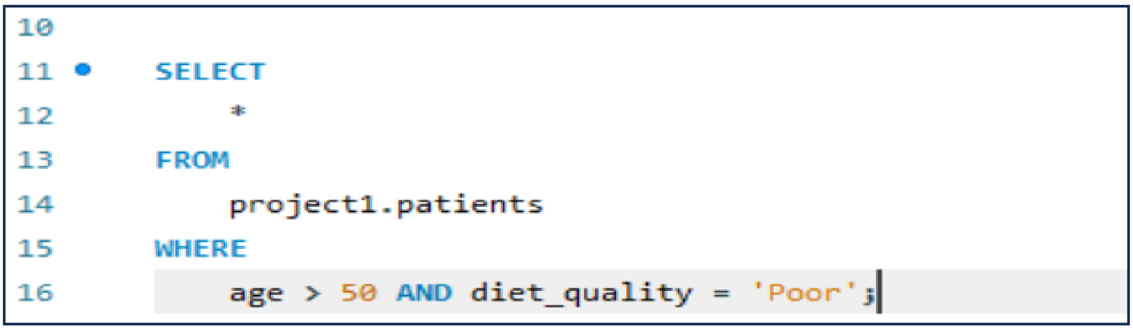
Project Title: Smoking and Cancer Risk Analysis using MY SQL
Name: Venkateswarlu Pujari
Date: 18th-04- 2025

Summary:

In this project, MySQL was used to analyse a healthcare dataset focused on understanding how smoking habits correlate with cancer risk. The data was organized across three relational tables—Patients, Habits, and Results—covering lifestyle, smoking behaviour, and medical outcomes. Through a series of SQL queries, the analysis identified high-risk groups, examined smoking patterns, calculated BMI trends, and evaluated cancer distribution. Key insights included patterns in cancer types by gender, risk levels based on smoking history, and cases of long-term smokers without diagnosis—offering valuable implications for preventive healthcare and research.

Query Analysis

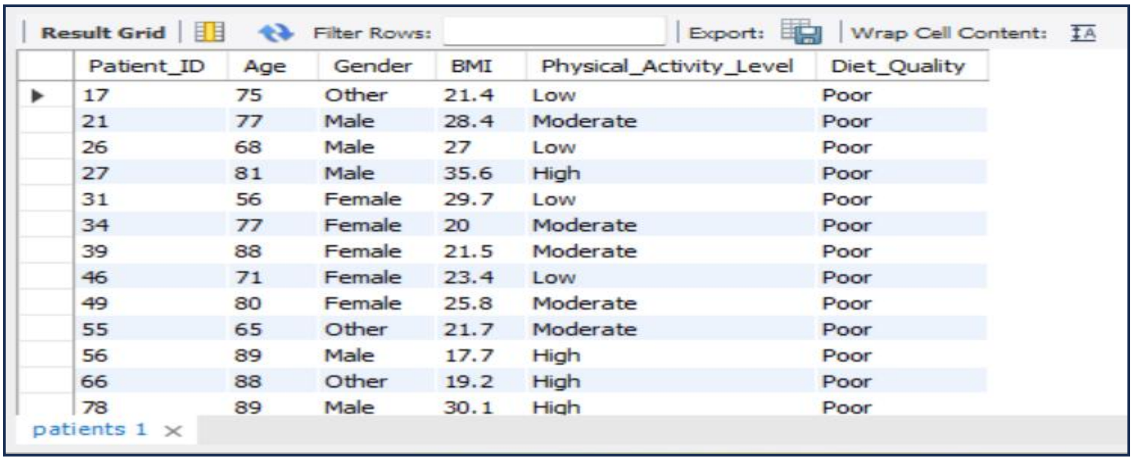
1. **Query:** List all patients who are older than 50 and have a poor diet.



```

SELECT
    *
FROM
    project1.patients
WHERE
    age > 50 AND diet_quality = 'Poor';
```

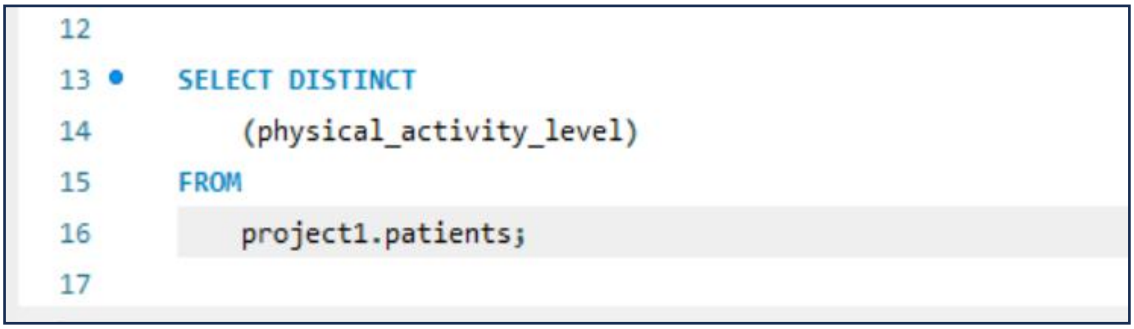
Output:



Patient_ID	Age	Gender	BMI	Physical_Activity_Level	Diet_Quality
17	75	Other	21.4	Low	Poor
21	77	Male	28.4	Moderate	Poor
26	68	Male	27	Low	Poor
27	81	Male	35.6	High	Poor
31	56	Female	29.7	Low	Poor
34	77	Female	20	Moderate	Poor
39	88	Female	21.5	Moderate	Poor
46	71	Female	23.4	Low	Poor
49	80	Female	25.8	Moderate	Poor
55	65	Other	21.7	Moderate	Poor
56	89	Male	17.7	High	Poor
66	88	Other	19.2	High	Poor
78	89	Male	30.1	High	Poor

Insight: These are high-risk patients based on age and dietary habits.

2. **Query:** Display the distinct physical activity levels recorded in the dataset

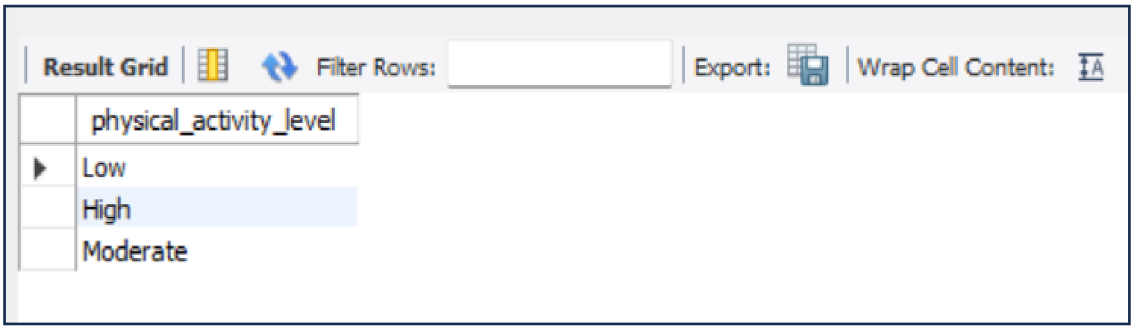


```

SELECT DISTINCT
    (physical_activity_level)
FROM
    project1.patients;

```

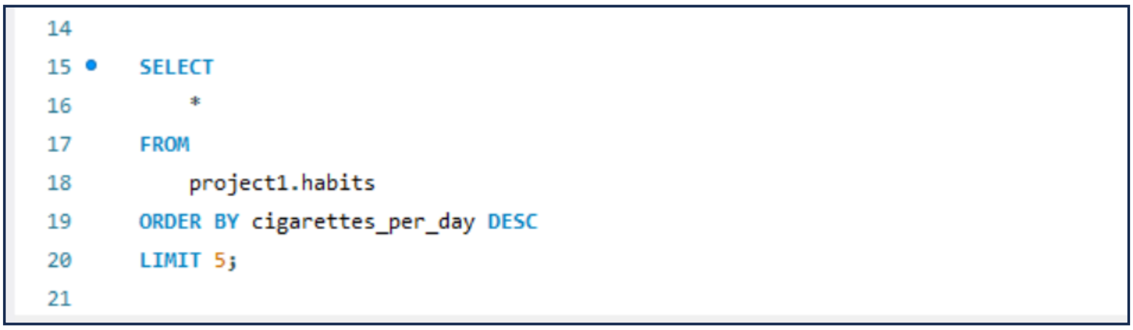
Output:



physical_activity_level
Low
High
Moderate

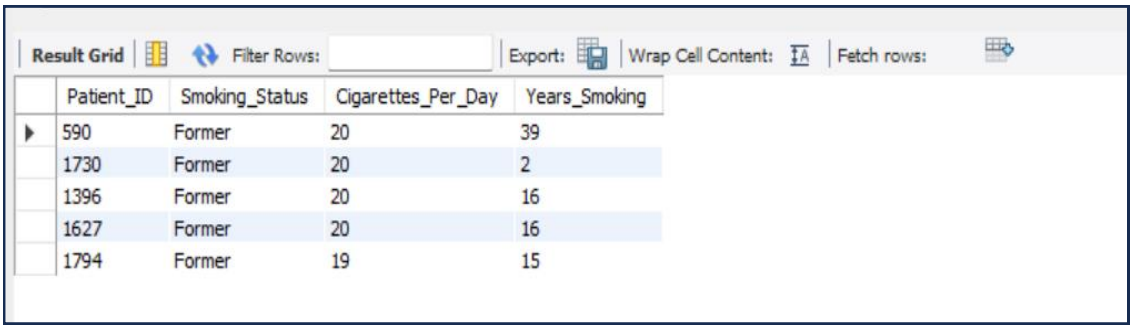
Insight: The dataset includes distinct Low, Moderate, and High activity levels.

3. **Query:** Show the top 5 patients with the highest number of cigarettes smoked per day



```
SELECT
    *
FROM
    project1.habits
ORDER BY cigarettes_per_day DESC
LIMIT 5;
```

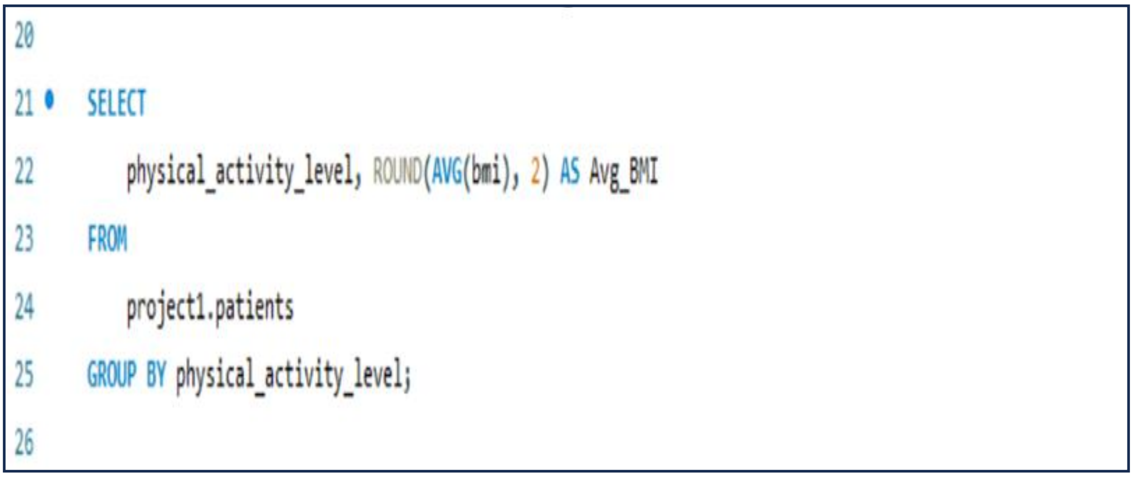
Output:



Patient_ID	Smoking_Status	Cigarettes_Per_Day	Years_Smoking
590	Former	20	39
1730	Former	20	2
1396	Former	20	16
1627	Former	20	16
1794	Former	19	15

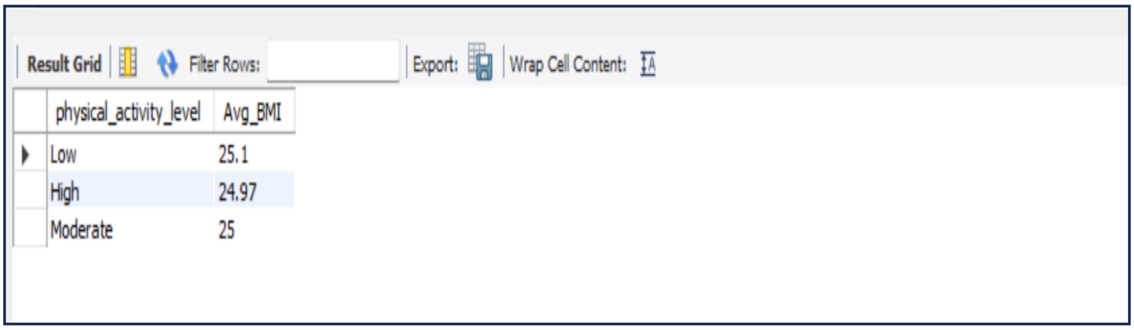
Insight: These patients have the highest daily tobacco exposure.

4. **Query:** Find the average BMI grouped by physical activity level.



```
SELECT
    physical_activity_level, ROUND(AVG(bmi), 2) AS Avg_BMI
FROM
    project1.patients
GROUP BY physical_activity_level;
```

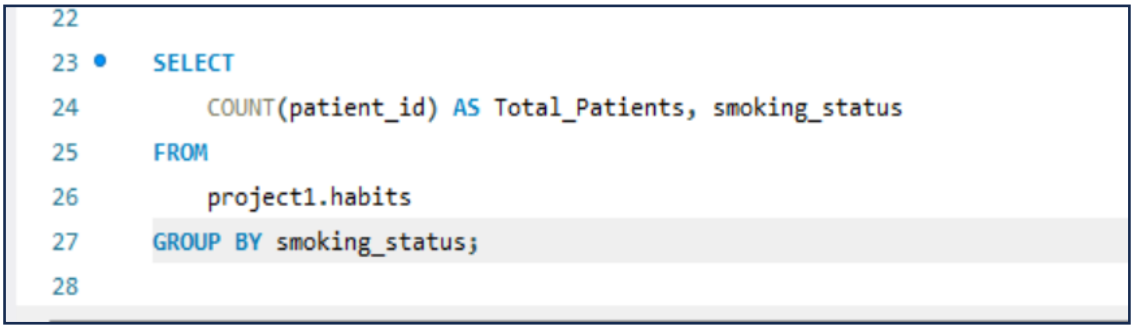
Output:



physical_activity_level	Avg_BMI
Low	25.1
High	24.97
Moderate	25

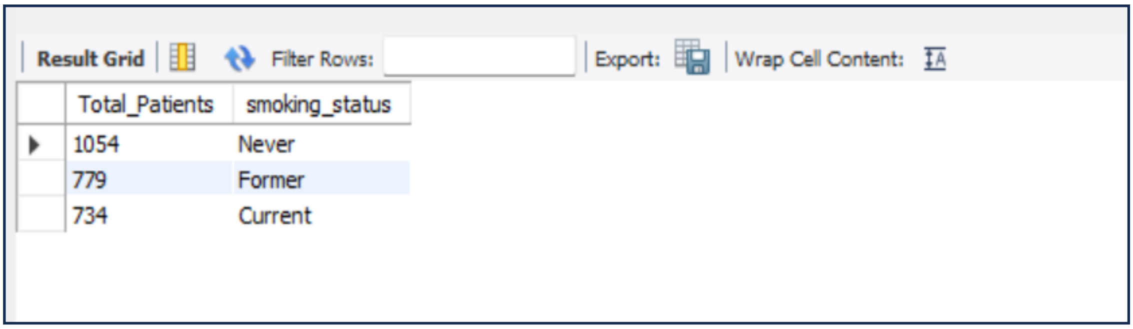
Insight: Patients with high physical activity levels tend to have lower BMI on average.

5. **Query:** Count how many patients fall into each smoking status category.



```
SELECT
    COUNT(patient_id) AS Total_Patients, smoking_status
FROM
    project1.habits
GROUP BY smoking_status;
```

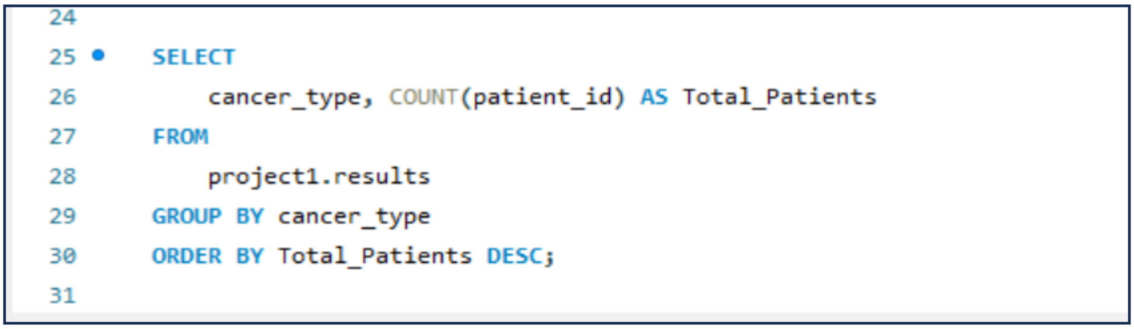
Output:



Total_Patients	smoking_status
1054	Never
779	Former
734	Current

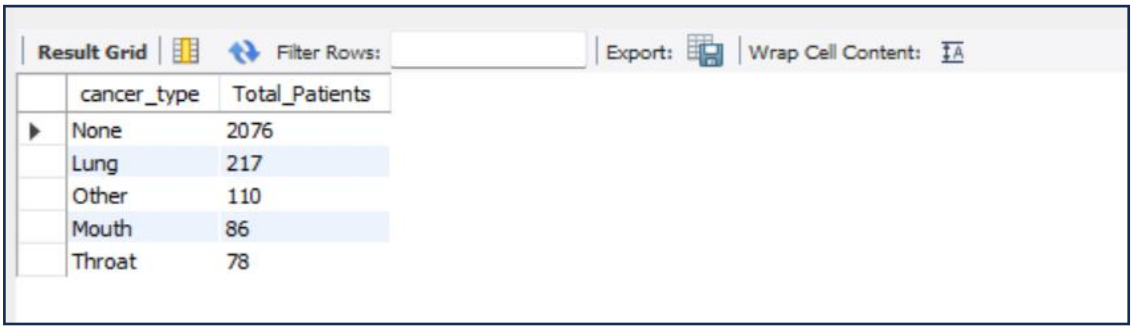
Insight: Most patients have never smoked. The remaining patients are more former smokers than current smokers.

6. **Query:** List the number of cancer patients per cancer type ordered by patient count



```
SELECT
    cancer_type, COUNT(patient_id) AS Total_Patients
FROM
    project1.results
GROUP BY cancer_type
ORDER BY Total_Patients DESC;
```

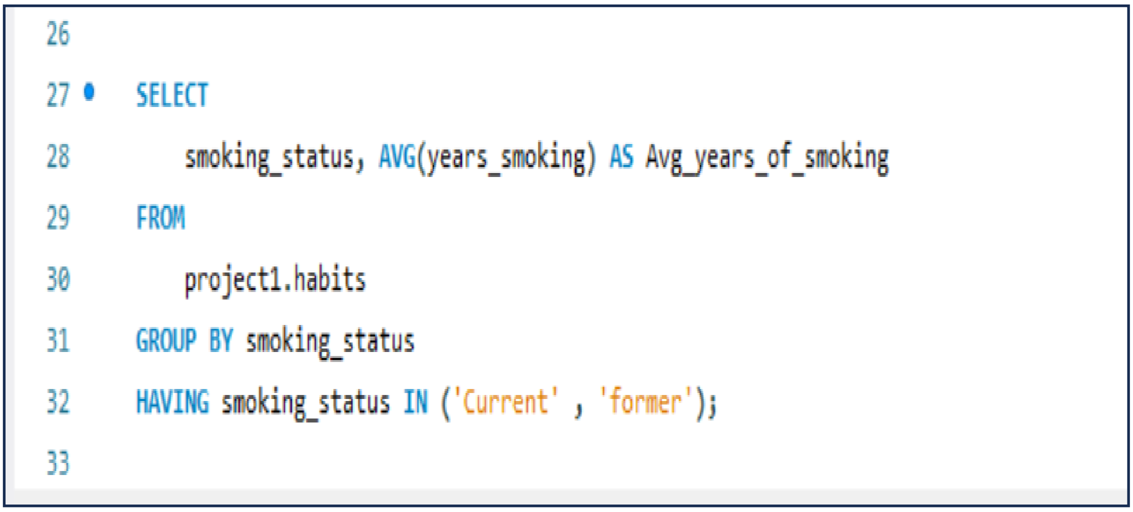
Output:



cancer_type	Total_Patients
None	2076
Lung	217
Other	110
Mouth	86
Throat	78

Insight: The highest concentration of patients falls under the "None" category.

7. **Query:** Find the average years of smoking for patients who are current and former smokers.

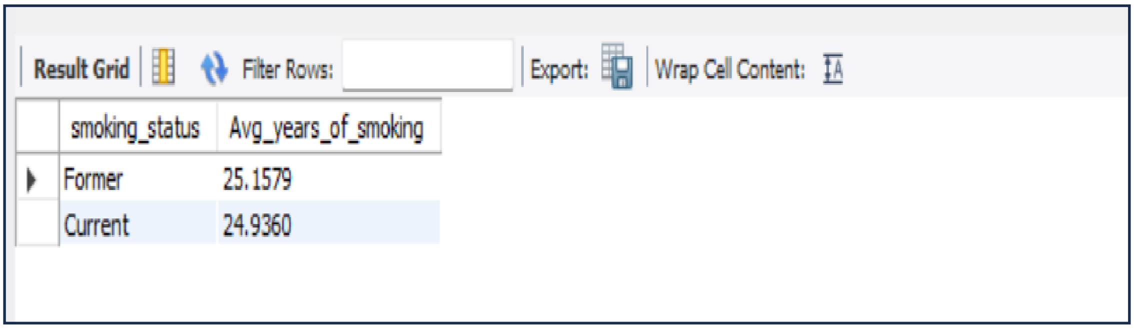


```sql

SELECT
    smoking_status, AVG(years_smoking) AS Avg_years_of_smoking
FROM
    project1.habits
GROUP BY smoking_status
HAVING smoking_status IN ('Current' , 'former');

```

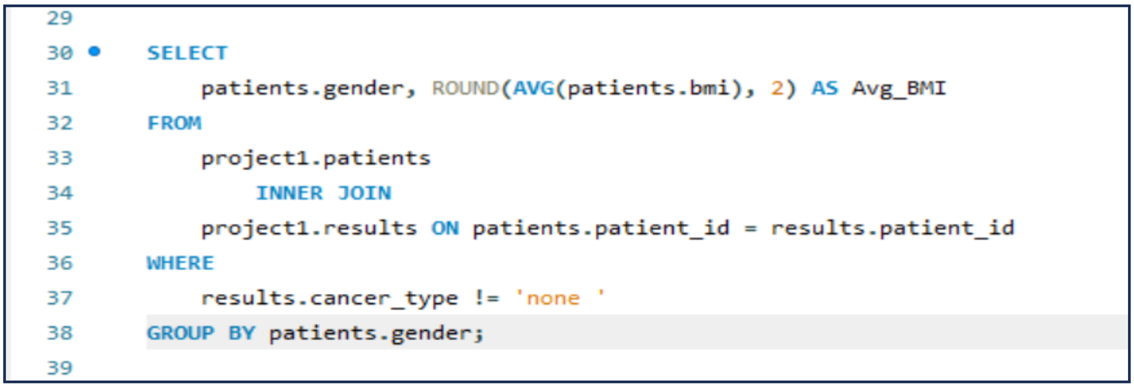
Output:



smoking_status	Avg_years_of_smoking
Former	25.1579
Current	24.9360

Insight: Former smokers have a slightly longer smoking history compared to current smokers

8. **Query:** Get the gender-wise average BMI of patients with cancer.

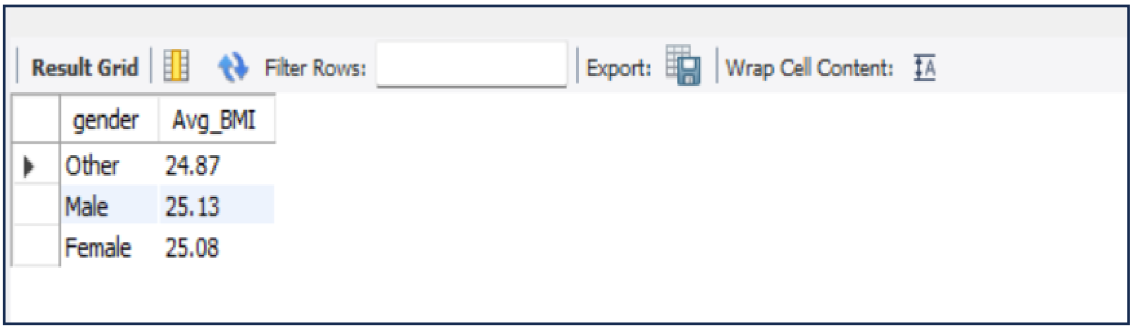


```sql

SELECT
    patients.gender, ROUND(AVG(patients.bmi), 2) AS Avg_BMI
FROM
    project1.patients
        INNER JOIN
    project1.results ON patients.patient_id = results.patient_id
WHERE
    results.cancer_type != 'none '
GROUP BY patients.gender;

```

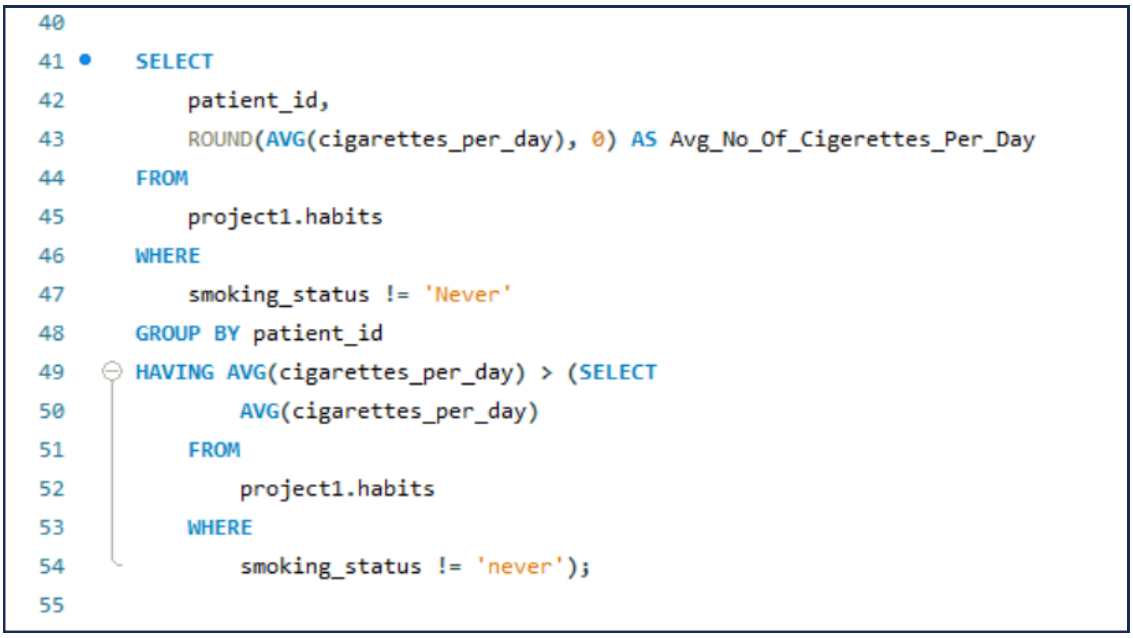
Output:



gender	Avg_BMI
Other	24.87
Male	25.13
Female	25.08

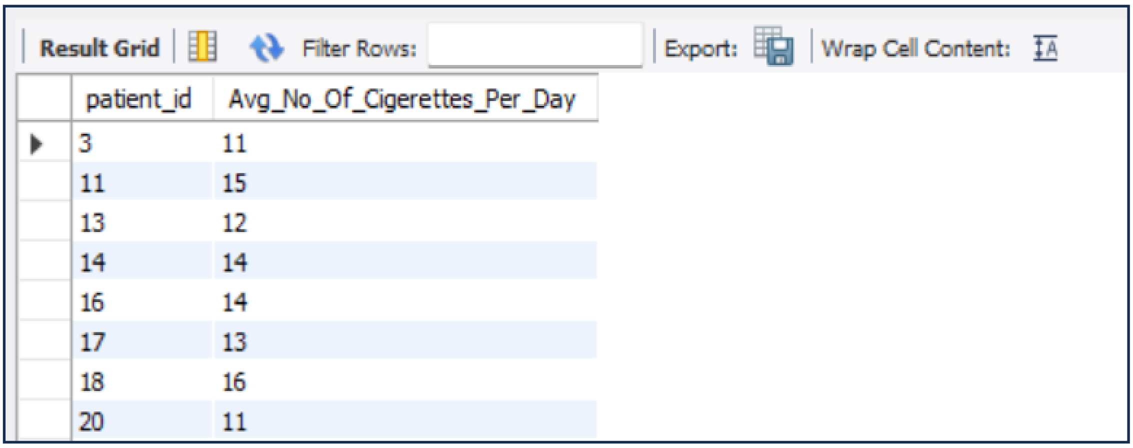
Insight: Males have the highest average BMI compared to females and other gender categories

9. **Query:** List patients who smoke more than the average number of cigarettes per day.



```
SELECT
    patient_id,
    ROUND(AVG(cigarettes_per_day), 0) AS Avg_No_Of_Cigerettes_Per_Day
FROM
    project1.habits
WHERE
    smoking_status != 'Never'
GROUP BY patient_id
HAVING AVG(cigarettes_per_day) > (SELECT
        AVG(cigarettes_per_day)
    FROM
        project1.habits
    WHERE
        smoking_status != 'never');
```

Output:

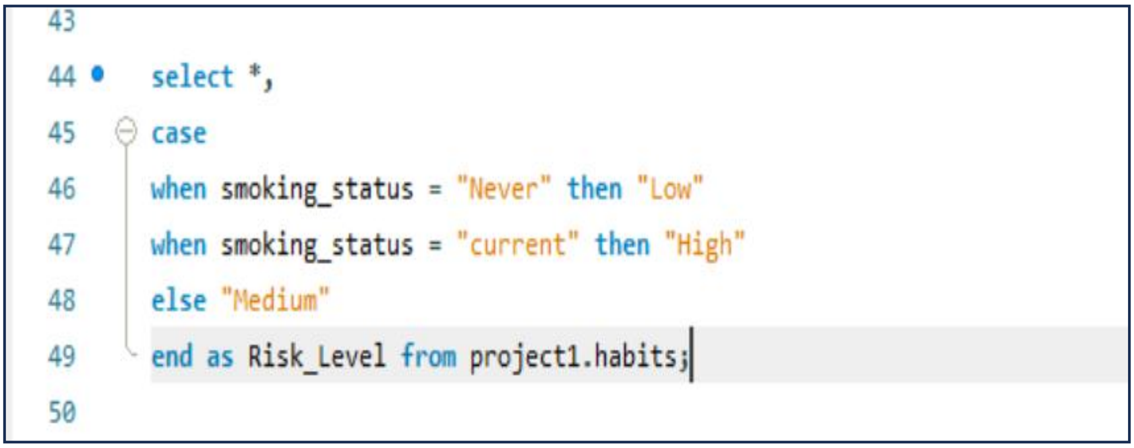


patient_id	Avg_No_Of_Cigerettes_Per_Day
3	11
11	15
13	12
14	14
16	14
17	13
18	16
20	11

Insight: This outcome shows the average number of cigarettes smoked per day by Patients

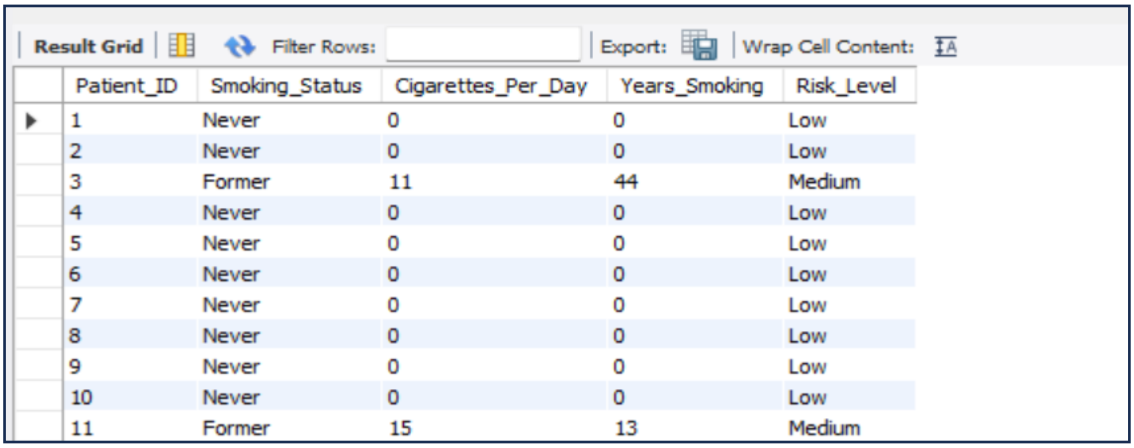
10. Query: Create a new column named Risk_Level using CASE:

- "High" if smoking status is 'Current'
- "Medium" if smoking status is 'Former'
- Else "Low"



```sql
select *,
case
when smoking_status = "Never" then "Low"
when smoking_status = "current" then "High"
else "Medium"
end as Risk_Level from project1.habits;
```

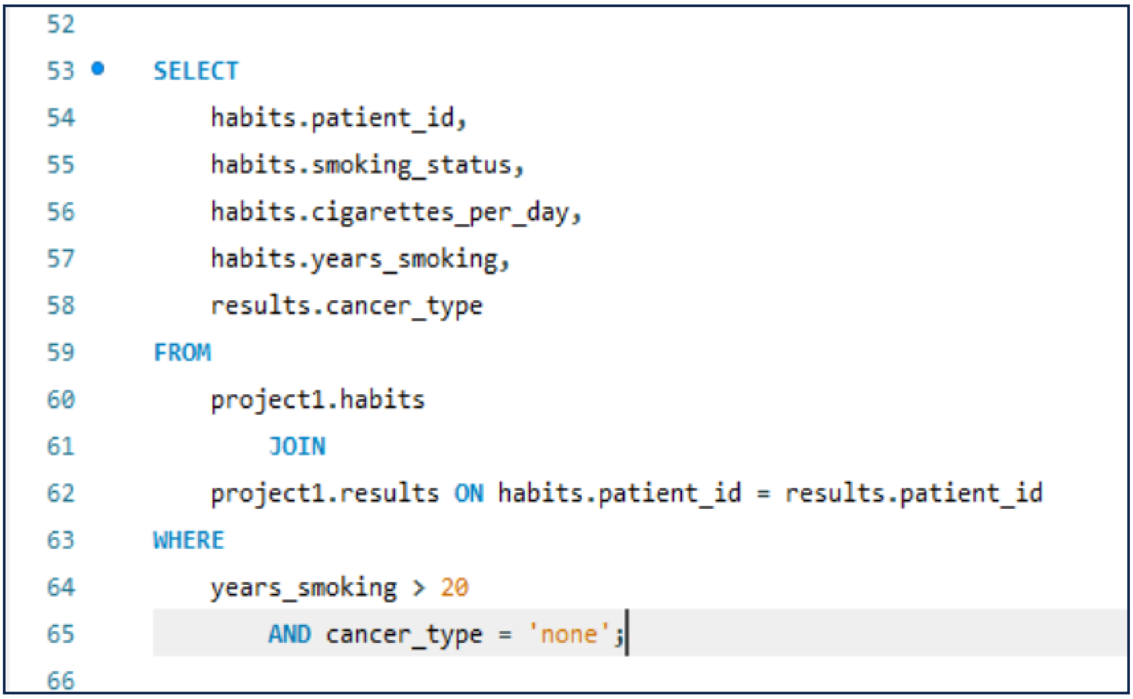
Output:



Patient_ID	Smoking_Status	Cigarettes_Per_Day	Years_Smoking	Risk_Level
1	Never	0	0	Low
2	Never	0	0	Low
3	Former	11	44	Medium
4	Never	0	0	Low
5	Never	0	0	Low
6	Never	0	0	Low
7	Never	0	0	Low
8	Never	0	0	Low
9	Never	0	0	Low
10	Never	0	0	Low
11	Former	15	13	Medium

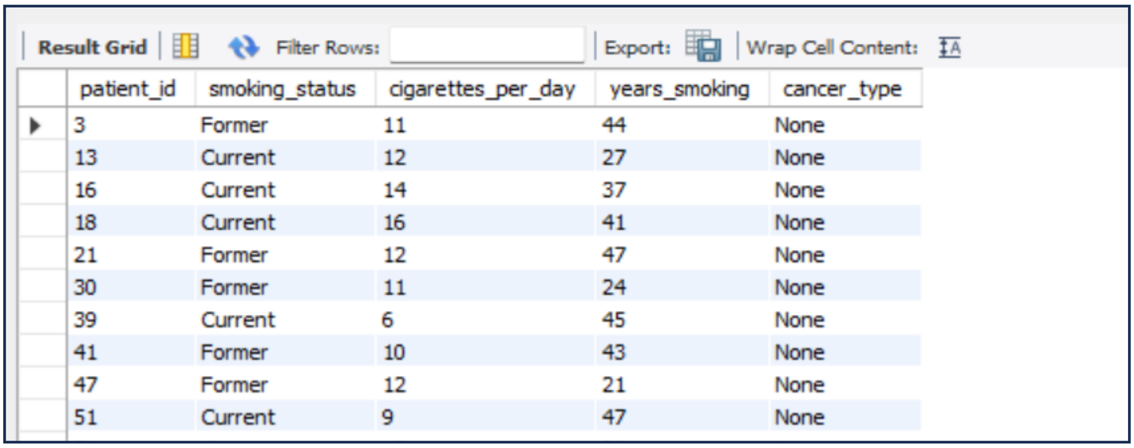
Insight: Patients who have never smoked exhibit a low risk level while former smokers have medium level and current smokers are high risk level

11. Query: Identify patients who have smoked for more than 20 years but have not been diagnosed with any cancer.



```sql
SELECT
    habits.patient_id,
    habits.smoking_status,
    habits.cigarettes_per_day,
    habits.years_smoking,
    results.cancer_type
FROM
    project1.habits
        JOIN
    project1.results ON habits.patient_id = results.patient_id
WHERE
    years_smoking > 20
        AND cancer_type = 'none';
```

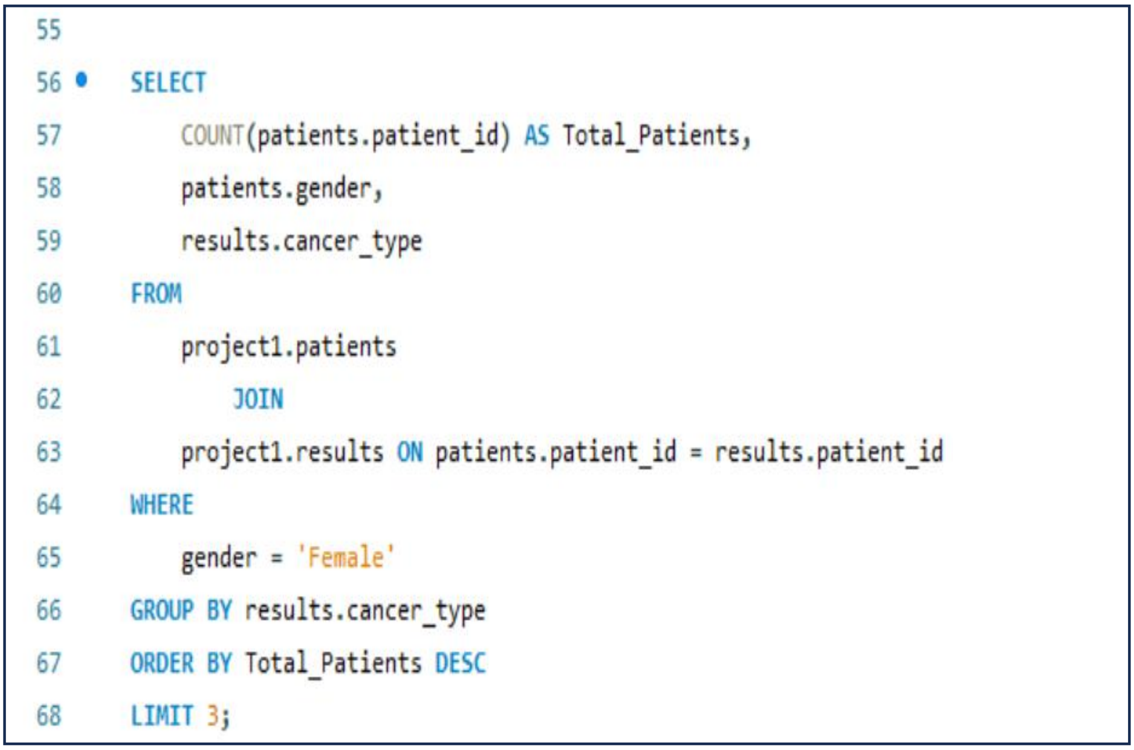
Output:



patient_id	smoking_status	cigarettes_per_day	years_smoking	cancer_type
3	Former	11	44	None
13	Current	12	27	None
16	Current	14	37	None
18	Current	16	41	None
21	Former	12	47	None
30	Former	11	24	None
39	Current	6	45	None
41	Former	10	43	None
47	Former	12	21	None
51	Current	9	47	None

Insight: This outcome shows patients who have not been diagnosed with cancer.

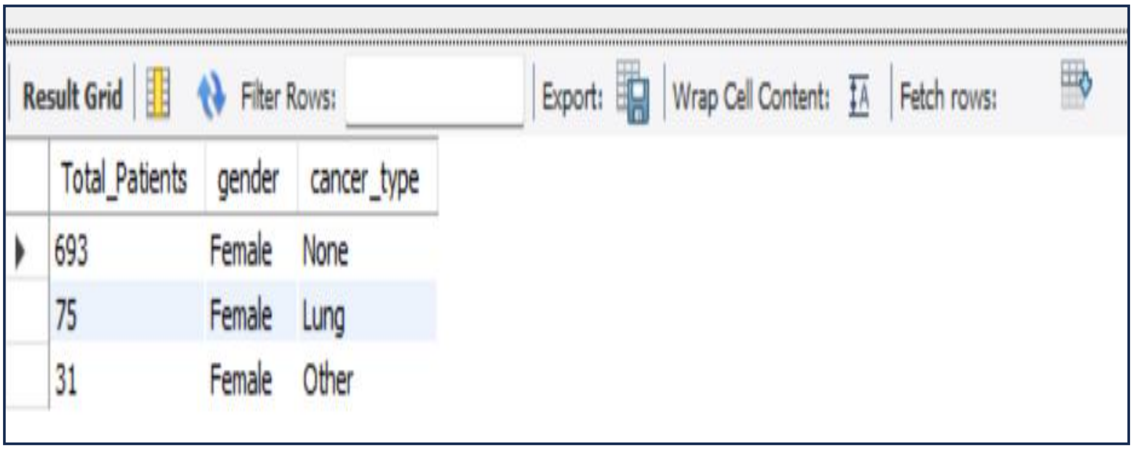
12. Query: Find the top 3 cancer types with the highest number of female patients.



```sql

SELECT
    COUNT(patients.patient_id) AS Total_Patients,
    patients.gender,
    results.cancer_type
FROM
    project1.patients
        JOIN
    project1.results ON patients.patient_id = results.patient_id
WHERE
    gender = 'Female'
GROUP BY results.cancer_type
ORDER BY Total_Patients DESC
LIMIT 3;
```

Output:



Total_Patients	gender	cancer_type
693	Female	None
75	Female	Lung
31	Female	Other

Insight: The majority of female patients in this dataset have no recorded cancer, while a smaller is diagnosed with lung cancer or other types.