Introduction

Thyroid disorders are a significant area of concern in endocrinology, influencing various aspects of health and well-being. The thyroid gland, responsible for producing hormones that regulate metabolism, can develop conditions such as goiter, hyperthyroidism, and thyroid cancer. Effective management and diagnosis of thyroid disorders rely heavily on analyzing comprehensive patient data, including demographics, medical history, and clinical outcomes. With a growing amount of data available, leveraging SQL for data analysis offers a powerful way to uncover insights and improve decision-making processes in thyroid care.

Main Problem

The central challenge in thyroid disorder management is the efficient and accurate analysis of patient data to inform diagnosis and treatment plans. The dataset provided includes various attributes such as age, gender, smoking history, thyroid function, and pathology results, which can impact the diagnosis and management of thyroid conditions. The main problems to address with SQL include:

- 1. Data Aggregation and Filtering: Efficiently aggregate and filter patient data to understand the distribution of different thyroid conditions and their outcomes.
 - Analyze the distribution of thyroid function types and their corresponding risk levels.
- 2. Risk Factor Analysis: Identify and analyze the key risk factors associated with various thyroid disorders.
- Determine how smoking history and thyroid function impact the risk level and recurrence rates.
- 3. Predictive Modeling: Utilize SQL queries to perform basic predictive analysis, such as identifying factors that influence treatment response or likelihood of recurrence.
- Query the data to find patterns that predict whether a patient is likely to experience recurrence.
- 4. Stage and Risk Analysis: Analyze the relationship between clinical stage and risk level, and how different factors influence these stages.
- Assess how physical examination findings and pathology results correlate with the stage of thyroid disease.
- 5. Outcome Assessment: Evaluate the effectiveness of different treatment responses and track patient outcomes based on their clinical and pathological data.
- Compare the effectiveness of treatment responses (e.g., 'Excellent' vs. 'Indeterminate') across different patient profiles.

SQL Queries for the Analysis

To address the outlined problems, you can use the following SQL queries as part of your portfolio project. These queries demonstrate how to aggregate, filter, and analyze the thyroid dataset.

1. Data Aggregation and Filtering

Query: Analyze the distribution of thyroid function types and their corresponding risk levels.

SELECT "Thyroid Function", Risk, **COUNT**(*) **AS** Number_of_Patients

FROM Thyroid_Diff

GROUP BY "Thyroid Function", Risk

ORDER BY Number_of_Patients **DESC**;

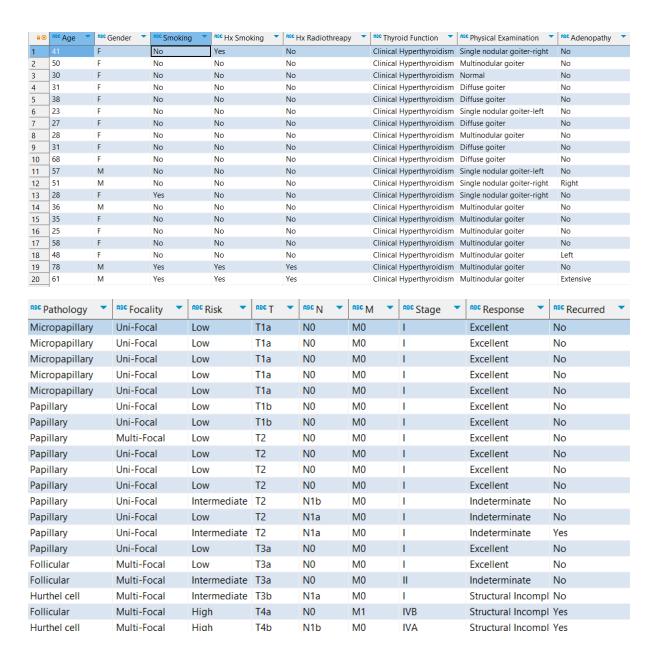
≙ ⊙	ABC Thyroid Function	Risk T	123 NUMBER_OF_PATIENTS
1	Euthyroid	Low	216
2	Euthyroid	Intermediate	88
3	Euthyroid	High	28
4	Clinical Hyperthyroidism	Low	14
5	Clinical Hypothyroidism	Low	9
6	Subclinical Hypothyroidism	Intermediate	7
7	Subclinical Hyperthyroidism	Low	5
8	Subclinical Hypothyroidism	Low	5
9	Clinical Hyperthyroidism	Intermediate	4
10	Clinical Hypothyroidism	Intermediate	3
11	Subclinical Hypothyroidism	High	2
12	Clinical Hyperthyroidism	High	2

Query: Filter the data to find patients with 'Clinical Hyperthyroidism'.

SELECT *

FROM Thyroid_Diff

WHERE "Thyroid Function" = 'Clinical Hyperthyroidism';



2. Risk Factor Analysis

Query: Determine how smoking history impacts the risk level.

SELECT Smoking, Risk, COUNT(*) AS Number_of_Patients
FROM Thyroid_Diff
GROUP BY Smoking, Risk
ORDER BY Number_of_Patients DESC;

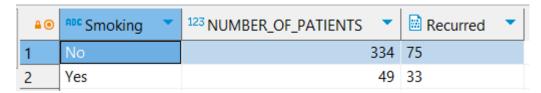
₽ ⊙	ABC Smoking T	Risk T	123 NUMBER_OF_PATIENTS
1	No	Low	234
2	No	Intermediate	86
3	Yes	High	18
4	Yes	Intermediate	16
5	Yes	Low	15
6	No	High	14

Query: Analyze the recurrence rates based on smoking history.

SELECT Smoking, COUNT(*) AS Number_of_Patients, SUM(CASE WHEN Recurred = 'Yes' THEN 1 ELSE 0 END) AS Recurred

FROM Thyroid_Diff

GROUP BY Smoking;



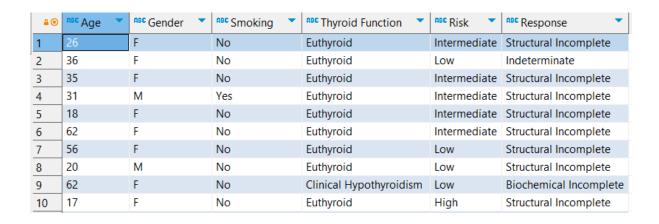
3. Predictive Modeling

Query: Identify factors that predict the likelihood of recurrence.

SELECT Age, Gender, Smoking, "Thyroid Function", Risk, Response

FROM Thyroid_Diff

WHERE Recurred = 'Yes';



Query: Predict treatment response based on initial clinical data.

SELECT Age, Gender, Smoking, "Thyroid Function", Risk, Pathology, Focality, Response **FROM** Thyroid_Diff

WHERE Risk = 'High';

₽ 💿	^{ABC} Age ▼	^{ABC} Gender ▼	ABC Smoking ~	ABC Thyroid Function	Risk T	^{ABC} Pathology ▼	^{ABC} Focality ▼	Response 🔻
1	17	F	No	Euthyroid	High	Papillary	Uni-Focal	Structural Incomplete
2	75	M	Yes	Euthyroid	High	Follicular	Multi-Focal	Structural Incomplete
3	34	F	No	Euthyroid	High	Papillary	Uni-Focal	Biochemical Incomplete
4	44	M	Yes	Euthyroid	High	Papillary	Multi-Focal	Structural Incomplete
5	38	F	No	Euthyroid	High	Papillary	Multi-Focal	Structural Incomplete
6	73	F	No	Euthyroid	High	Papillary	Multi-Focal	Structural Incomplete
7	35	М	Yes	Euthyroid	High	Papillary	Multi-Focal	Structural Incomplete
8	26	F	Yes	Euthyroid	High	Hurthel cell	Multi-Focal	Structural Incomplete
9	53	F	No	Euthyroid	High	Papillary	Uni-Focal	Structural Incomplete
10	35	F	No	Euthyroid	High	Papillary	Multi-Focal	Structural Incomplete
11	49	M	No	Euthyroid	High	Papillary	Multi-Focal	Structural Incomplete
12	34	F	No	Euthyroid	High	Papillary	Uni-Focal	Structural Incomplete
13	80	F	Yes	Euthyroid	High	Papillary	Uni-Focal	Structural Incomplete
14	67	F	No	Euthyroid	High	Papillary	Multi-Focal	Structural Incomplete
15	68	F	No	Euthyroid	High	Papillary	Multi-Focal	Structural Incomplete

4. Stage and Risk Analysis

Query: Assess the relationship between clinical stage and risk level.

SELECT Stage, Risk, COUNT(*) AS Number_of_Patients

FROM Thyroid_Diff

GROUP BY Stage, Risk;

4 ⊚	^{ABC} Stage ▼	ABC Risk	123 NUMBER_OF_PATIENTS
1	L	Low	248
2	I	Intermediate	77
3	II	Intermediate	24
4	I	High	8
5	IVB	High	11
6	II	Low	1
7	II	High	7
8	III	Intermediate	1
9	III	High	3
10	IVA	High	3

Query: Analyze how physical examination findings correlate with disease stage.

SELECT Physical_Examination, Stage, COUNT(*) AS Number_of_Patients FROM Thyroid_Diff

GROUP BY Physical_Examination, Stage;

≙ ⊚	ABC Physical Examination	^{ABC} Stage ▼	123 NUMBER_OF_PATIENTS
1	Single nodular goiter-left	I	76
2	Multinodular goiter	L	115
3	Single nodular goiter-right	L	128
4	Normal	I	7
5	Diffuse goiter	L	7
6	Multinodular goiter	II	14
7	Single nodular goiter-right	II	9
8	Single nodular goiter-left	IVB	3
9	Single nodular goiter-left	II	9
10	Single nodular goiter-left	III	1
11	Single nodular goiter-right	IVB	2
12	Multinodular goiter	III	2
13	Single nodular goiter-right	III	1
14	Multinodular goiter	IVB	6
15	Multinodular goiter	IVA	3

5. Outcome Assessment

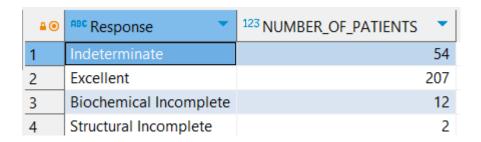
Query: Evaluate the effectiveness of different treatment responses.

SELECT Response, COUNT(*) AS Number_of_Patients

FROM Thyroid_Diff

WHERE Recurred = 'No'

GROUP BY Response;



Query: Track patient outcomes based on clinical and pathological data.

SELECT Response, Pathology, COUNT(*) AS Number_of_Patients

FROM Thyroid_Diff

GROUP BY Response, Pathology

ORDER BY Number_of_Patients DESC;

₽ ⊙	Response 🔻	Pathology •	123 NUMBER_OF_PATIENTS
1	Excellent	Papillary	149
2	Structural Incomplete	Papillary	75
3	Indeterminate	Papillary	44
4	Excellent	Micropapillary	43
5	Biochemical Incomplete	Papillary	19
6	Excellent	Follicular	9
7	Structural Incomplete	Follicular	9
8	Excellent	Hurthel cell	7
9	Structural Incomplete	Hurthel cell	7
10	Indeterminate	Hurthel cell	6
11	Indeterminate	Follicular	6
12	Indeterminate	Micropapillary	5
13	Biochemical Incomplete	Follicular	4

Conclusion

Using these complex SQL queries, you can extract valuable insights from the thyroid dataset, enabling a more nuanced understanding of the factors affecting thyroid disorders. These advanced analyses help in predicting outcomes, assessing treatment efficacy, and identifying trends, ultimately contributing to more informed clinical decision-making and personalized patient care.