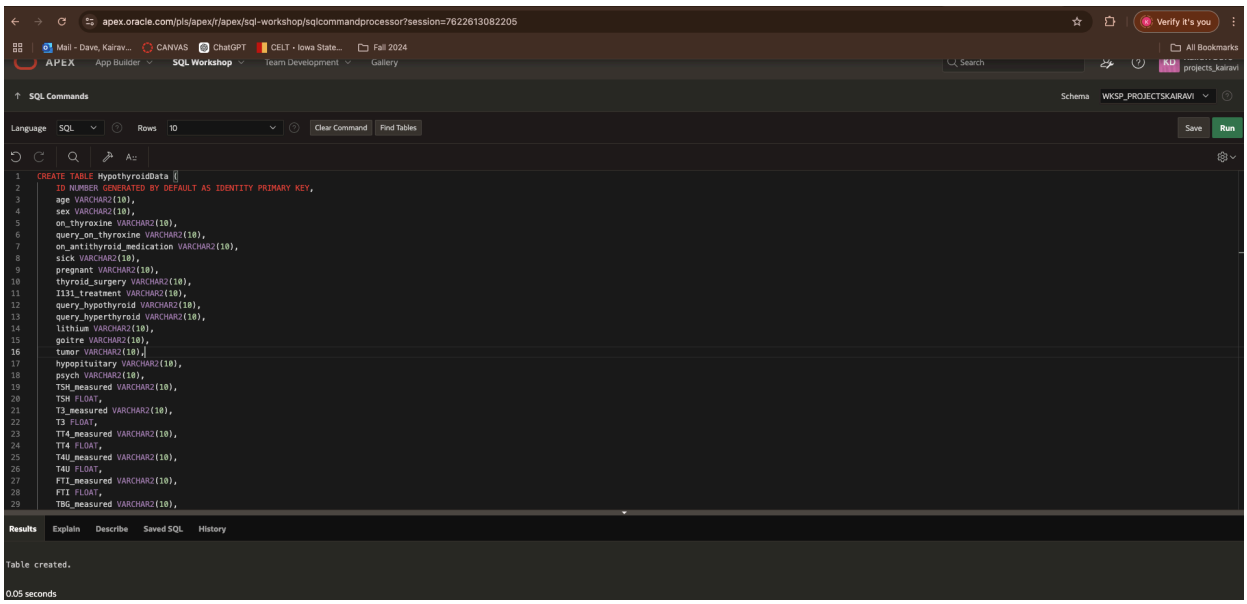
Data Cleaning and Preprocessing using SQL

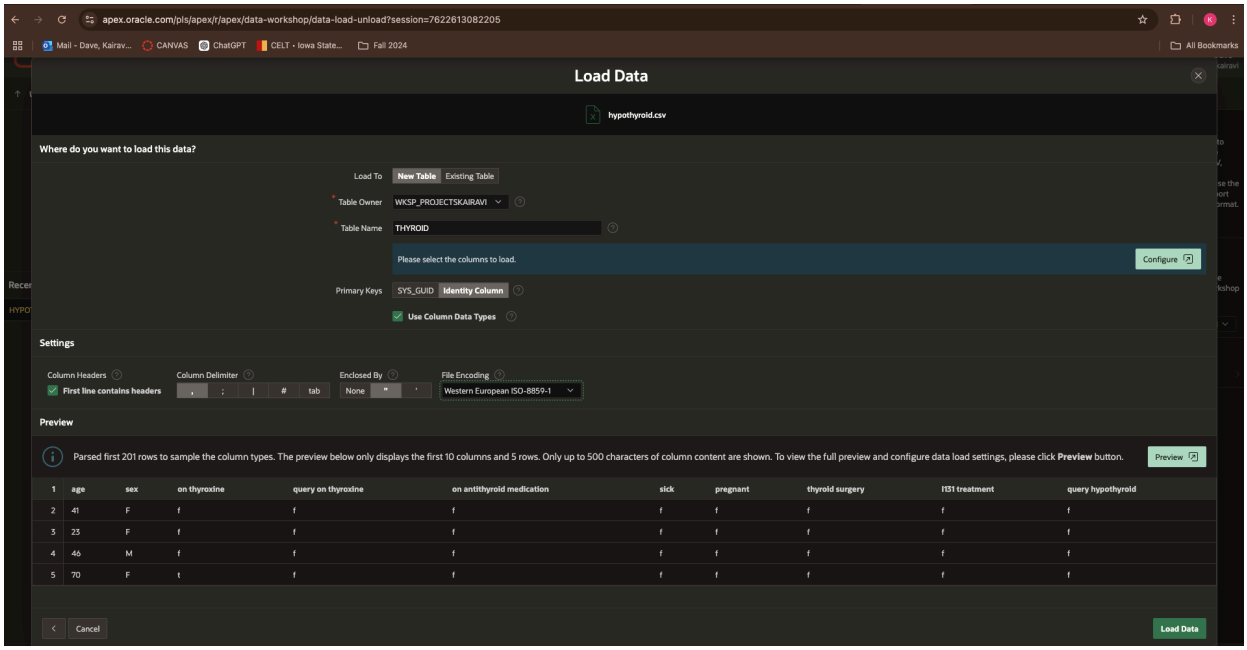
Dataset:
https://www.kaggle.com/code/yasserhessein/thyroid-disease-detection-using-deep-learning

Using APEX Oracle - SQL Commands




```
CREATE TABLE HypothyroidData (
   ID NUMBER GENERATED BY DEFAULT AS IDENTITY PRIMARY KEY,
   age VARCHAR2(10),
   sex VARCHAR2(10),
   on_thyroxine VARCHAR2(10),
   query_on_thyroxine VARCHAR2(10),
   on_antithyroid_medication VARCHAR2(10),
   sick VARCHAR2(10),
   pregnant VARCHAR2(10),
   thyroid_surgery VARCHAR2(10),
   I131_treatment VARCHAR2(10),
   query_hypothyroid VARCHAR2(10),
   query_hyperthyroid VARCHAR2(10),
   lithium VARCHAR2(10),
   goitre VARCHAR2(10),
   tumor VARCHAR2(10),
   hypopituitary VARCHAR2(10),
   psych VARCHAR2(10),
```

```
    TSH_measured VARCHAR2(10),
    TSH FLOAT,
    T3_measured VARCHAR2(10),
    T3 FLOAT,
    TT4_measured VARCHAR2(10),
    TT4 FLOAT,
    T4U_measured VARCHAR2(10),
    T4U FLOAT,
    FTI_measured VARCHAR2(10),
    FTI FLOAT,
    TBG_measured VARCHAR2(10),
    TBG FLOAT,
    referral_source VARCHAR2(50),
    binaryClass VARCHAR2(10)
);
```



Display 10 rows :

```
SELECT * FROM THYROID
FETCH FIRST 10 ROWS ONLY;
```

Language SQL Rows 10 Clear Command Find Tables Save Run

```
SELECT * FROM THYROID
FETCH FIRST 10 ROWS ONLY;
```

Results Explain Describe Saved SQL History

ID	AGE	SEX	ON_THYROXINE	QUERY_ON_THYROXINE	ON_ANTITHYROID_MEDICATION	SICK	PREGNANT	THYROID_SURGERY	I131_TREATMENT	QUERY_HYPOTHYROID	QUERY_HYPERTHYROID	LITHIUM	GOITRE	TUMOR	HYPOPITUITARY	PSYCH
1	41	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f
2	23	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f
3	46	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f
4	70	F	t	f	f	f	f	f	f	f	f	f	f	f	f	f
5	70	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f
6	18	F	t	f	f	f	f	f	f	f	f	f	f	f	f	f
7	59	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f
8	80	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f
9	66	F	f	f	f	f	f	f	f	f	f	f	f	t	f	f
10	68	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f

10 rows returned in 0.01 seconds Download

Step 1: Data Cleaning

1.1 Handling Missing Values or Invalid Values

```
SELECT
   COUNT(*) AS TotalRows,
   COUNT(CASE WHEN AGE IS NULL THEN 1 END) AS Missing_Age,
   COUNT(CASE WHEN TSH IS NULL THEN 1 END) AS Missing_TSH,
   COUNT(CASE WHEN T3 IS NULL THEN 1 END) AS Missing_T3,
   COUNT(CASE WHEN TT4 IS NULL THEN 1 END) AS Missing_TT4
FROM THYROID;
```

Language SQL Rows 10 Clear Command Find Tables Save Run

```
SELECT
    COUNT(*) AS TotalRows,
    COUNT(CASE WHEN AGE IS NULL THEN 1 END) AS Missing_Age,
    COUNT(CASE WHEN TSH IS NULL THEN 1 END) AS Missing_TSH,
    COUNT(CASE WHEN T3 IS NULL THEN 1 END) AS Missing_T3,
    COUNT(CASE WHEN TT4 IS NULL THEN 1 END) AS Missing_TT4
FROM THYROID;
```

Results Explain Describe Saved SQL History

TOTALROWS	MISSING_AGE	MISSING_TSH	MISSING_T3	MISSING_TT4
3771	0	369	769	231

1 rows returned in 0.01 seconds Download

1.2 Fixing Missing/ Invalid Values

```
UPDATE THYROID
SET TSH = (SELECT AVG(TSH) FROM THYROID WHERE TSH IS NOT NULL)
WHERE TSH IS NULL;
```

```sql
UPDATE THYROID
SET TSH = (SELECT AVG(TSH) FROM THYROID WHERE TSH IS NOT NULL)
WHERE TSH IS NULL;
```

Results | Explain | Describe | Saved SQL | History

369 row(s) updated.

0.02 seconds

Same for T3 and TT4

Verify the cleaning

```sql
SELECT
    COUNT(*) AS TotalRows,
    COUNT(CASE WHEN AGE IS NULL THEN 1 END) AS Missing_Age,
    COUNT(CASE WHEN TSH IS NULL THEN 1 END) AS Missing_TSH,
    COUNT(CASE WHEN T3 IS NULL THEN 1 END) AS Missing_T3,
    COUNT(CASE WHEN TT4 IS NULL THEN 1 END) AS Missing_TT4
FROM THYROID;
```

Results | Explain | Describe | Saved SQL | History

TOTALROWS	MISSING_AGE	MISSING_TSH	MISSING_T3	MISSING_TT4
3771	0	0	0	0

1 rows returned in 0.01 seconds Download

Step 2: Exploratory Data Analysis (EDA)

2.1 Understanding the data distribution

What is the distribution of Thyroid cases (hypothyroid vs non-hypothyroid)?

SELECT binaryClass, COUNT(*) AS TotalCases
FROM THYROID
GROUP BY binaryClass;

```sql
SELECT binaryClass, COUNT(*) AS TotalCases
FROM THYROID
GROUP BY binaryClass;
```

Results | Explain | Describe | Saved SQL | History

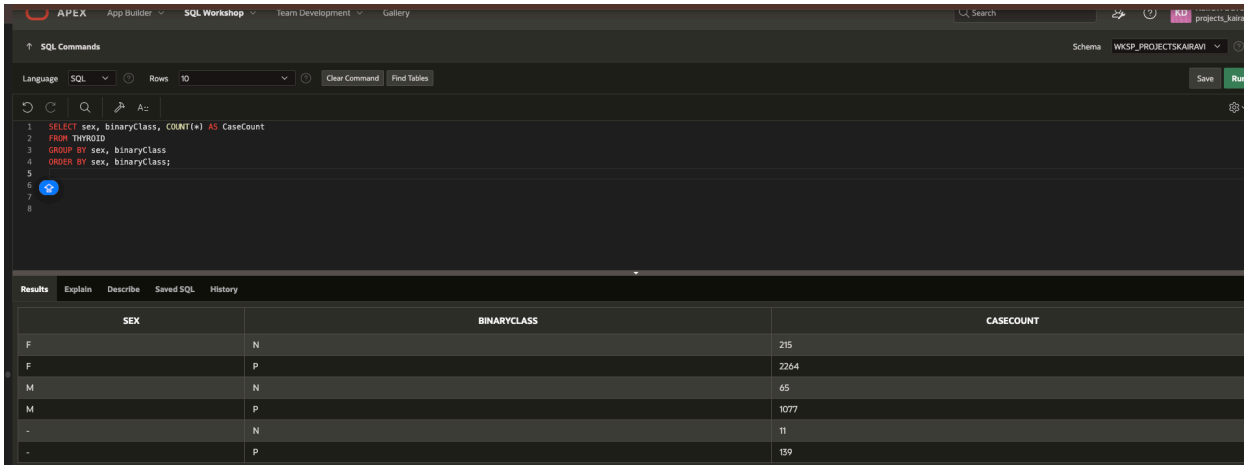
BINARYCLASS	TOTALCASES
P	3480
N	291

2 rows returned in 0.01 seconds Download

2.2 Distribution of Thyroid Cases by Gender

How does thyroid disease vary between males and females ?

```
SELECT sex, binaryClass, COUNT(*) AS CaseCount
FROM THYROID
GROUP BY sex, binaryClass
ORDER BY sex, binaryClass;
```

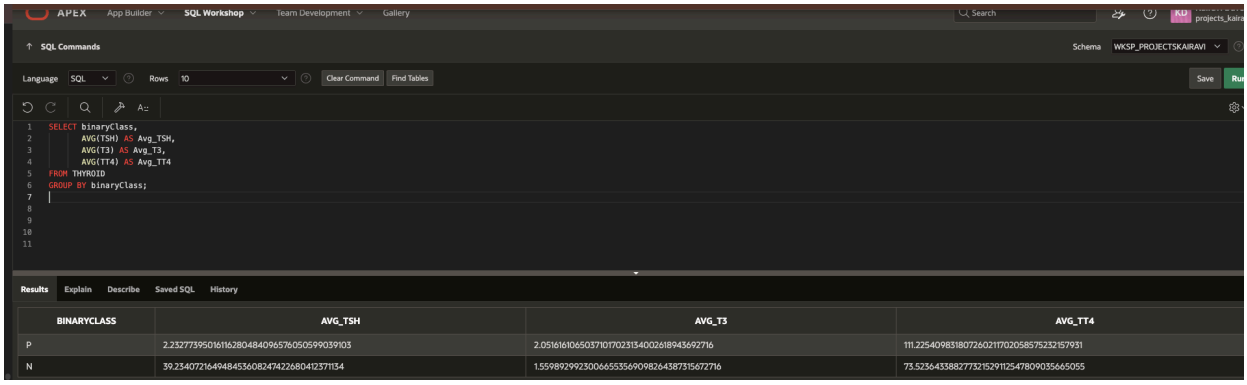


SEX	BINARYCLASS	CASECOUNT
F	N	215
F	P	2264
M	N	65
M	P	1077
-	N	11
-	P	139

2.3 Average TSH, T3, AND TT4 Levels

What are the average TSH, T3, TT4 levels for hypothyroid vs non-hypothyroid patients?

```
SELECT binaryClass,
    AVG(TSH) AS Avg_TSH,
    AVG(T3) AS Avg_T3,
    AVG(TT4) AS Avg_TT4
FROM THYROID
GROUP BY binaryClass;
```



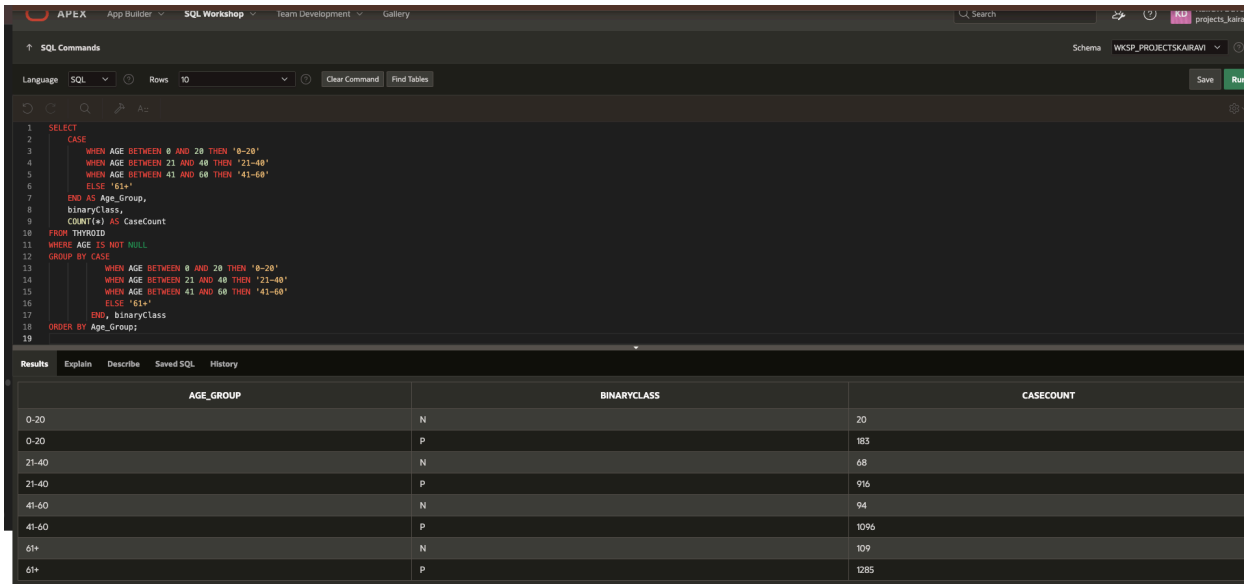
BINARYCLASS	AVG_TSH	AVG_T3	AVG_TT4
P	2.232773950161162804840965760505990391 03	2.05161610650371017023134002618943692716	111.22540983180726021170205857523215793 1
N	39.23407216494845360824742268041237113 4	1.55989299230066553569098264387315672716	73.52364338827732152911254780903566505 5

2.4 Age Group Analysis

How are hypothyroid cases distributed across age groups?

SELECT

```
    CASE
        WHEN AGE BETWEEN 0 AND 20 THEN '0-20'
        WHEN AGE BETWEEN 21 AND 40 THEN '21-40'
        WHEN AGE BETWEEN 41 AND 60 THEN '41-60'
        ELSE '61+'
    END AS Age_Group,
    binaryClass,
    COUNT(*) AS CaseCount
FROM THYROID
WHERE AGE IS NOT NULL
GROUP BY CASE
            WHEN AGE BETWEEN 0 AND 20 THEN '0-20'
            WHEN AGE BETWEEN 21 AND 40 THEN '21-40'
            WHEN AGE BETWEEN 41 AND 60 THEN '41-60'
            ELSE '61+'
        END, binaryClass
ORDER BY Age_Group;
```



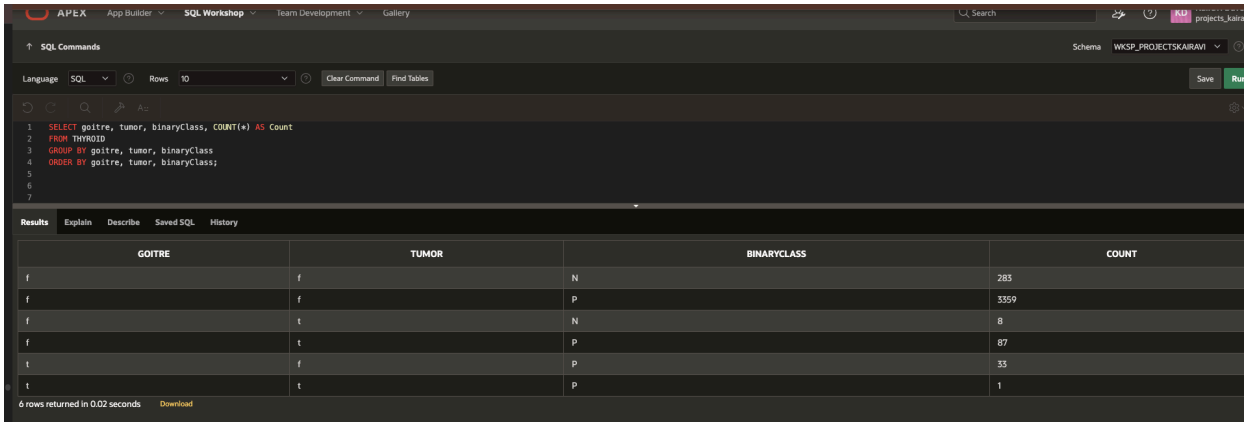
AGE_GROUP	BINARYCLASS	CASECOUNT
0-20	N	20
0-20	P	183
21-40	N	68
21-40	P	916
41-60	N	94
41-60	P	1096
61+	N	109
61+	P	1285

2.5 Impact of Goitre and Tumors

How does the presence of Goitre and Tumor relate to thyroid case?

```
SELECT goitre, tumor, binaryClass, COUNT(*) AS Count
FROM THYROID
GROUP BY goitre, tumor, binaryClass
```

ORDER BY goitre, tumor, binaryClass;



GOITRE	TUMOR	BINARYCLASS	COUNT
f	f	N	283
f	f	P	3359
f	t	N	8
f	t	P	87
t	f	P	33
t	t	P	1

2.6 Relationships between Pregnancy and Hypothyroidism

Are pregnant women more prone to hypothyroidism?

SELECT pregnant, binaryClass, COUNT(*) AS CaseCount
FROM THYROID
GROUP BY pregnant, binaryClass
ORDER BY pregnant, binaryClass;

PREGNANT	BINARYCLASS	CASECOUNT
f	N	291
f	P	3427
t	P	53