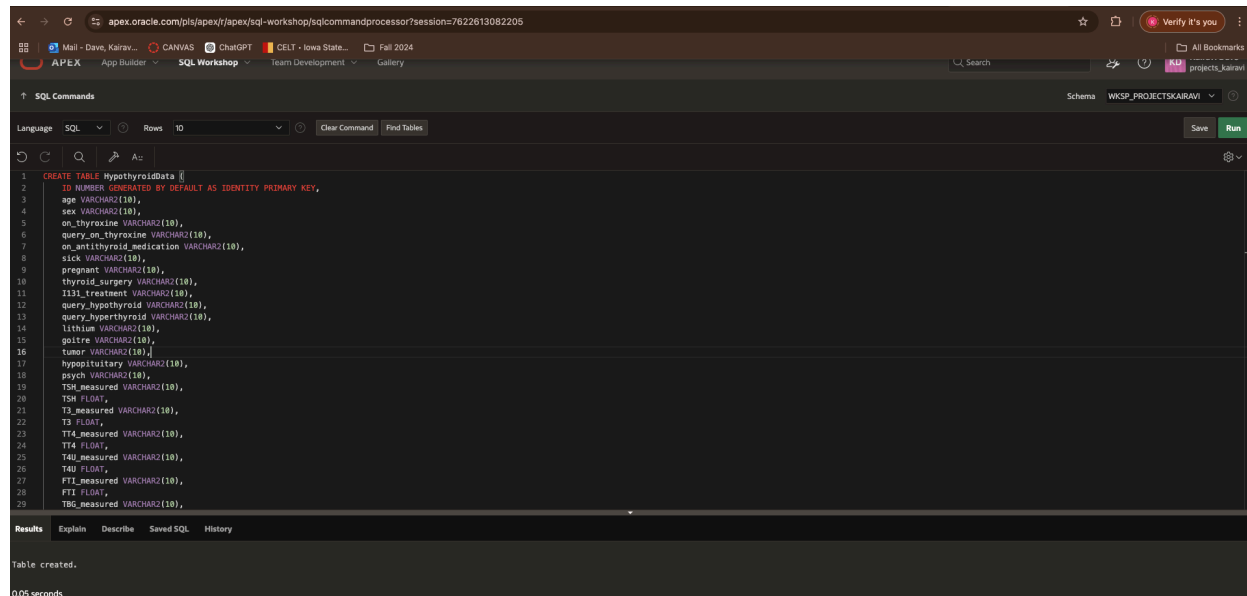


# Data Cleaning and Preprocessing using SQL

## Dataset:

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

## Using APEX Oracle - SQL Commands

The screenshot shows the APEX Oracle SQL Workshop interface. The top navigation bar includes links to Mail, Canvas, ChatGPT, and CELT. The main area is titled 'SQL Commands' and shows a successful execution of a CREATE TABLE statement. The table 'HypoThyroidData' has been created with various columns including age, sex, on\_thyroxine, query\_on\_thyroxine, on\_antithyroid\_medication, sick, pregnant, thyroid\_surgery, l131\_treatment, query\_hypothyroid, query\_hyperthyroid, lithium, goitre, tumor, hypopituitary, psych, TSH\_measured, T3\_measured, T3, T4\_measured, T4, FTI\_measured, FTI, and TBG\_measured. The results pane at the bottom shows 'Table created.' and '0.05 seconds'.

```
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),  
  l131_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),  
  T3_measured VARCHAR2(10),  
  T3 FLOAT,  
  T4_measured VARCHAR2(10),  
  T4 FLOAT,  
  FTI_measured VARCHAR2(10),  
  FTI FLOAT,  
  TBG_measured 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)
);

```

Where do you want to load this data?

Load To: **New Table** Existing Table

Table Owner: WKSP\_PROJECTSKARAVI

Table Name: THYROID

Please select the columns to load. [Configure](#)

Primary Keys: SYS\_GUID, Identity Column

☒ Use Column Data Types

Settings

Column Headers: ☒ First line contains headers

Column Delimiter: comma

Enclosed By: None

File Encoding: Western European ISO-8859-1

Preview

Parsed first 201 rows to sample the column types. The preview below only displays the first 10 columns and 5 rows. Only up to 500 characters of column content are shown. To view the full preview and configure data load settings, please click [Preview](#) button.

	age	sex	on thyroxine	query on thyroxine	on antithyroid medication	sick	pregnant	thyroid surgery	T3T4 treatment	query hypothyroid
1	41	F	f	f	f	f	f	f	f	f
2	23	F	f	f	f	f	f	f	f	f
3	46	M	f	f	f	f	f	f	f	f
4	70	F	t	f	f	f	f	f	f	f
5										

[Cancel](#) [Load Data](#)

Display 10 rows :

```

SELECT * FROM THYROID
FETCH FIRST 10 ROWS ONLY;

```

Language: SQL Rows: 10 Clear Command Find Tables Save Run

```

1 SELECT * FROM THYROID;
2 FETCH FIRST 10 ROWS ONLY;
3

```

ID	AGE	SEX	ON_THYROXINE	QUERY_ON_THYROXINE	ON_ANTITHYROID_MEDICATION	SICK	PREGNANT	THYROID_SURGERY	H3L_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

```

1 SELECT
2   COUNT(*) AS TotalRows,
3   COUNT(CASE WHEN AGE IS NULL THEN 1 END) AS Missing_Age,
4   COUNT(CASE WHEN TSH IS NULL THEN 1 END) AS Missing_TSH,
5   COUNT(CASE WHEN T3 IS NULL THEN 1 END) AS Missing_T3,
6   COUNT(CASE WHEN TT4 IS NULL THEN 1 END) AS Missing_TT4
7 FROM THYROID;
8
9
10
11

```

TOTALROWS	MISSING_AGE	MISSING_TSH	MISSING_T3	MISSING_TT4
5771	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;

```

1 UPDATE THYROID
2 SET TSH = (SELECT Avg(TSH) FROM THYROID WHERE TSH IS NOT NULL )
3 WHERE TSH IS NULL;

```

Results

369 row(s) updated.  
0.02 seconds

Same for T3 and TT4  
Verify the cleaning

```

1 SELECT
2 COUNT(*) AS TotalRows,
3 COUNT(CASE WHEN AGE IS NULL THEN 1 END) AS Missing_Age,
4 COUNT(CASE WHEN TSH IS NULL THEN 1 END) AS Missing_TSH,
5 COUNT(CASE WHEN T3 IS NULL THEN 1 END) AS Missing_T3,
6 COUNT(CASE WHEN TT4 IS NULL THEN 1 END) AS Missing_TT4
7 FROM THYROID;

```

Results

TOTALROWS	MISSING_AGE	MISSING_TSH	MISSING_T3	MISSING_TT4
5771	0	0	0	0

1 rows returned in 0.01 seconds

## 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;

```

```

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

```

Results

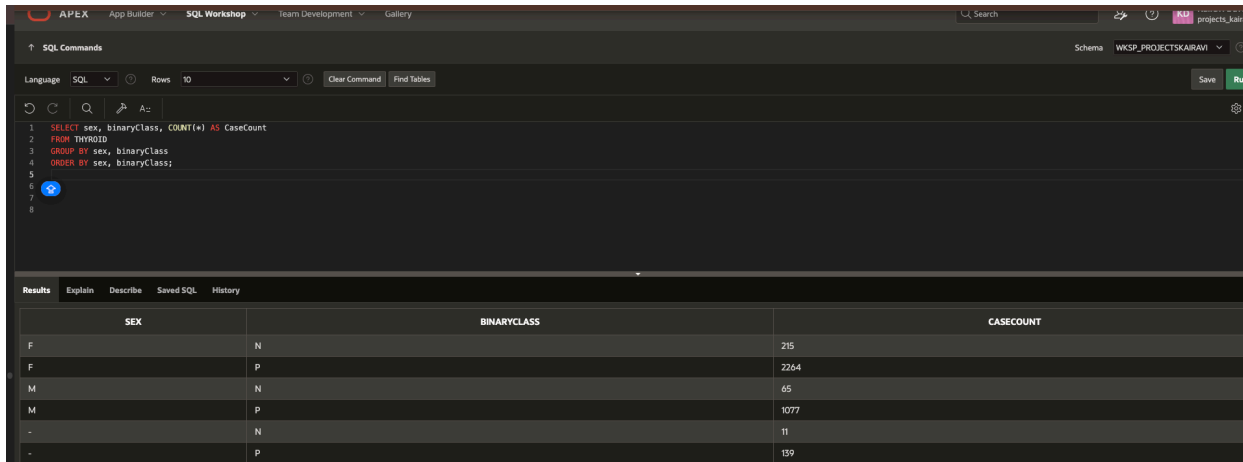
BINARYCLASS	TOTALCASES
P	3480
N	291

2 rows returned in 0.01 seconds

## 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;
```



The screenshot shows the APEX SQL Workshop interface. The SQL command area contains the following query:

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

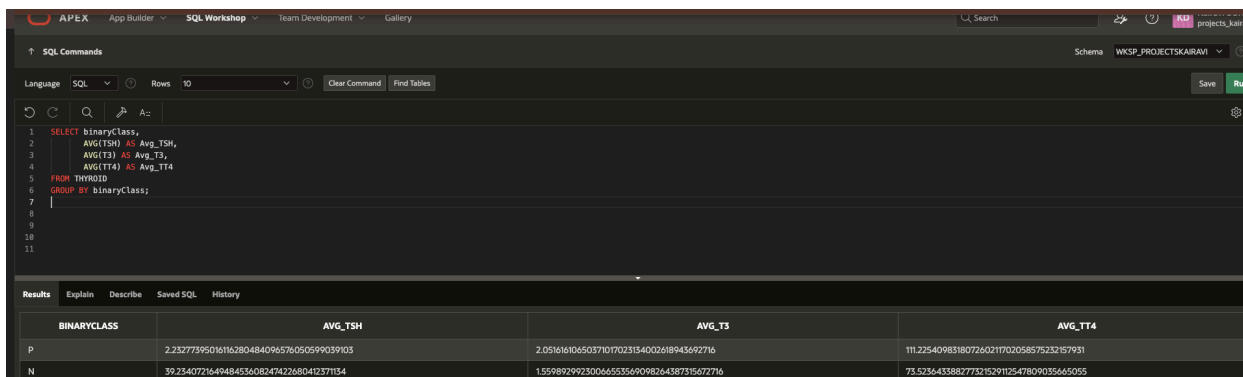
The Results tab is active, displaying a table with the following data:

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;
```



The screenshot shows the APEX SQL Workshop interface. The SQL command area contains the following query:

```
1 SELECT binaryClass,
2        AVG(TSH) AS Avg_TSH,
3        AVG(T3) AS Avg_T3,
4        AVG(TT4) AS Avg_TT4
5 FROM THYROID
6 GROUP BY binaryClass;
```

The Results tab is active, displaying a table with the following data:

BINARYCLASS	AVG_TSH	AVG_T3	AVG_TT4
P	2.23277395016162804840965760505990339103	2.05161610650371017023134002618943692716	111.225409831807260211702058575232157931
N	39.234072164948453608247422680412371134	1.5598929230066553569098264387315672716	73.523643388277321529112547809035665055

## 2.4 Age Group Analysis

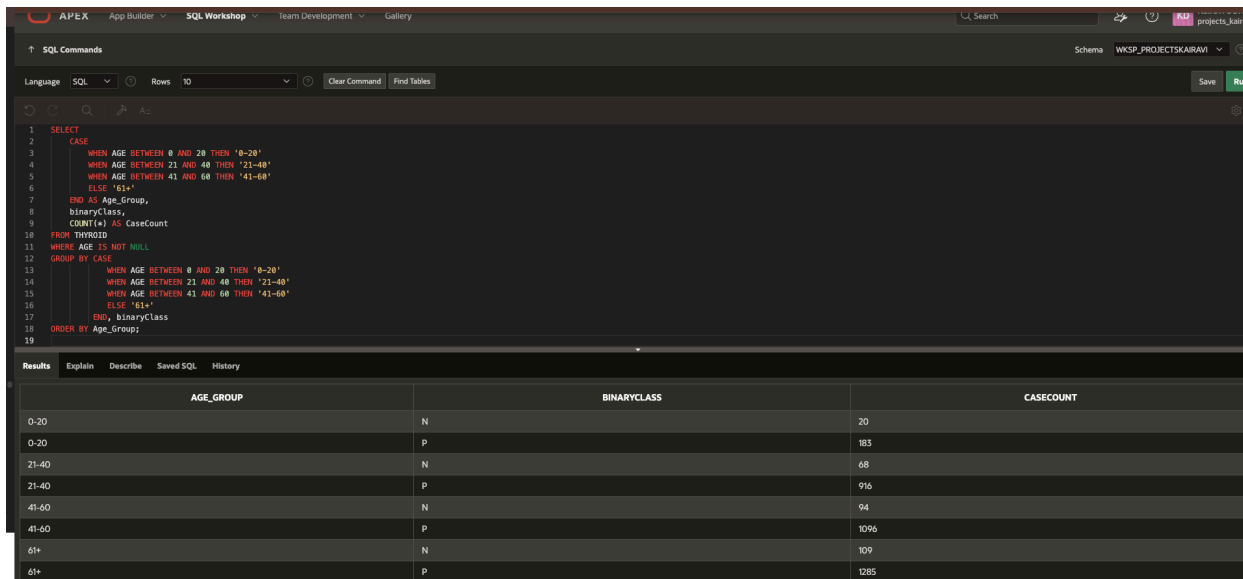
How are hypothyroid cases distributed across age groups?

```
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;

```



The screenshot shows the Oracle APEX SQL Workshop interface. The SQL command window contains the query from the previous block. The results window displays the output of the query, which is a table with three columns: AGE\_GROUP, BINARYCLASS, and CASECOUNT. The results are grouped by AGE\_GROUP and BINARYCLASS.

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

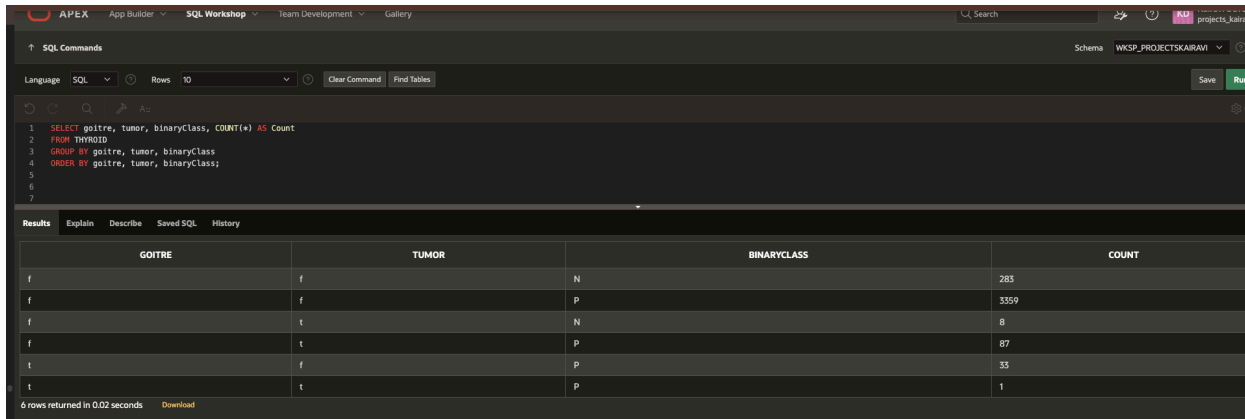
How does the presence of Goitre and Tumor relate to thyroid case?

```

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

```

ORDER BY goitre, tumor, binaryClass;



The screenshot shows the APEX SQL Workshop interface. The SQL command area contains the following query:

```
1 SELECT goitre, tumor, binaryClass, COUNT(*) AS Count
2 FROM THYROID
3 GROUP BY goitre, tumor, binaryClass
4 ORDER BY goitre, tumor, binaryClass;
5
6
7
```

The Results tab is selected, displaying a table with 6 rows and 4 columns: GOITRE, TUMOR, BINARYCLASS, and COUNT. The data is as follows:

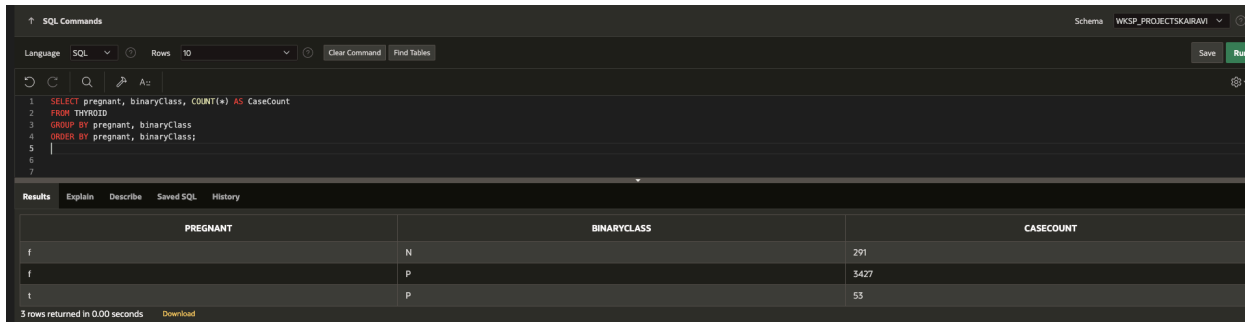
GOITRE	TUMOR	BINARYCLASS	COUNT
f	f	N	283
f	f	P	3359
f	t	N	8
f	t	P	87
t	f	P	33
t	t	P	1

6 rows returned in 0.02 seconds

## 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;
```



The screenshot shows the APEX SQL Workshop interface. The SQL command area contains the following query:

```
1 SELECT pregnant, binaryClass, COUNT(*) AS CaseCount
2 FROM THYROID
3 GROUP BY pregnant, binaryClass
4 ORDER BY pregnant, binaryClass;
5
6
7
```

The Results tab is selected, displaying a table with 3 rows and 3 columns: PREGNANT, BINARYCLASS, and CASECOUNT. The data is as follows:

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

3 rows returned in 0.00 seconds