Univariate and Bivariate Analysis – Capstone project

Univariate

1. **Mean, Median, Mode**

	Age	TSH_Level	T3_Level	T4_Level	Nodule_Size
Mean	51.918497	5.045102	2.001727	8.246204	2.503403
Median	52.0	5.04	2.0	8.24	2.51
Mode	72	9.9	1.27	10.67	0.69

Feature	Normal Range	Dataset Median	Dataset Mode
TSH Level (mIU/L)	0.4 – 4.0	5.04	9.9
T3 Level (nmol/L)	1.2 – 2.8	2	1.27
T4 Level (µg/dL)	4.5 – 12.0	8.24	10.67
Nodule Size (cm)	<1 cm benign; >1 cm clinically significant	2.51	0.69

- The Mean (average) Age around 51.9 implies most of the patients belong to around 50 age .
- The Mean (average) TSH_Level is around 5 implies average TSH_level is 5.
- The Mean (average) T3_level is around 2 implies most of the T3_Level is 2.
- The Mean (average) T4_Level is around 8.2 implies most of the T4_Level is 8.2.
- The Mean (average is around 2.5 implies most of the patients Nodule_Size is 2.5.
- The Median (middle value) Age is 52 years implies majority of patients fall in the middle-aged group.
- The Median (middle value) TSH_Level is 5.04 which means that half of the patients have TSH levels below 5.04, and the other half have levels above it.
- The Median (middle value) T3_level is around 2 implies majority of patients have T3 hormone levels within a normal range.
- The Median (middle value) T4_Level is around 8.24 majority of patients have T4 levels within a normal range.
- The Median (middle value) Nodule_Size is around 2.51 implies most patients have moderately sized thyroid nodules.
- The Mode (most common) Age is 72, means many patients belong to around 72 years.
- The Mode (most common) TSH_Level is 9.9, which indicates that many individuals have relatively **high TSH levels** compared to the normal range.
- The Mode (most common) T3_Level is 1.27, means that a large number of patients in the dataset have **normal or slightly lower T3 levels**.
- The Mode (most common) T4_Level is 10.67, indicates that a considerable portion of patients may exhibit **slightly elevated T4 activity**.

- ❖ The Mode (most common) Nodule size is 0.69cm, implying that small nodules are the most common among patients.

1. **Percentile Report**

	Age	TSH_Level	T3_Level	T4_Level	Nodule_Size
Mean	51.918497	5.045102	2.001727	8.246204	2.503403
Median	52.0	5.04	2.0	8.24	2.51
Mode	72	9.9	1.27	10.67	0.69
Q1:25%	33.0	2.57	1.25	6.37	1.25
Q2:50%	52.0	5.04	2.0	8.24	2.51
Q3:75%	71.0	7.52	2.75	10.12	3.76
Q4:100%	89.0	10.0	3.5	12.0	5.0

Age:

- ➢ The 25th percentile (Q1) is 33 — means 25% of patients are below 33 years.
- ➢ (Q2) 50% of the Age column is 52 — means half of the patients are below 52, indicating that most patients are middle-aged.
- ➢ From Q1 to Q2 there is a 57.6% increase showing that a majority of patients are in the middle-aged range (33–52 years).
- ➢ The 75th percentile (Q3) is 71 — 75% of patients are below 71 years.
- ➢ There is about a 36.5% rise in age from the median (52 years) to the 75th percentile (71 years), showing that the dataset includes a some portion of older patients.
- ➢ The maximum age (Q4) is 89 — indicating a few elderly patients in the dataset.

TSH_Level (Thyroid Stimulating Hormone):

- ➢ The 25th percentile (Q1) is 2.57, meaning 25% of patients have TSH levels below 2.57.
- ➢ The 50th percentile (Q2) or median is 5.04, meaning half of the patients have TSH levels below 5.04.
- ➢ From Q1 to Q2, there is a 96% increase, indicating that some patients have normal readings and some patients have elevated levels.
- ➢ The 75th percentile (Q3) is 7.52, meaning 75% of patients have TSH levels below 7.52.
- ➢ There is about a 49% rise in TSH levels from Q2 (5.04) to the 75th percentile Q3 (7.52), showing that some patients have higher-than-normal values.
- ➢ The maximum TSH value (Q4) is 10.0, indicating a few patients with very high TSH levels, possibly representing hypothyroid cases.

T3_Level (Triiodothyronine):

- The 25th percentile (Q1) is 1.25, meaning 25% of patients have T3 levels below 1.25.
- The 50th percentile (Q2) or median is 2.00, meaning half of the patients have T3 levels below 2.00.
- From Q1 to Q2, there is a 60% increase, indicating that most patients' T3 levels are within the normal range.
- The 75th percentile (Q3) is 2.75, meaning 75% of patients have T3 levels below 2.75.
- There is a 37.5% increase from the median to the 75th percentile, showing a few patients with relatively higher T3 values.
- The maximum T3 value (Q4) is 3.5, indicating that some patients may have slightly elevated T3 levels suggestive of hyperthyroid tendencies.

T4_Level (Thyroxine):

- The 25th percentile (Q1) is 6.37, meaning 25% of patients have T4 levels below 6.37.
- The 50th percentile (Q2) or median is 8.24, meaning half of the patients have T4 levels below 8.24.
- From Q1 to Q2, there is a 29.4% increase, showing a moderate rise in T4 levels within the normal range.
- The 75th percentile (Q3) is 10.12, meaning 75% of patients have T4 levels below 10.12.
- There is a 22.8% increase from the median to the 75th percentile, showing that some patients have slightly elevated T4 levels.
- The maximum T4 value (Q4) is 12.0, indicating a few patients with high T4 readings, reflecting mild hyperthyroidism.

Nodule_Size (cm):

- The 25th percentile (Q1) is 1.25, means 25% of patients have nodules smaller than 1.25 cm.
- The 50th percentile (Q2) or median is 2.51, meaning half of the patients have nodules smaller than 2.51 cm.
- From Q1 to Q2, there is a 100.8% increase, indicating that nodule size roughly doubles between these percentiles.
- The 75th percentile (Q3) is 3.76, meaning 75% of patients have nodules below 3.76 cm.
- There is a 49.8% increase from the median to the 75th percentile, showing some patients with moderately large nodules.
- The maximum nodule size (Q4) is 5.0, indicating the presence of a few large nodules in the dataset.

2. Distribution plots

Distribution plots help you understand how each numeric feature is spread – like normal, right skewed, Left skewed, Uniform. They also help in visually spot outliers and understand data concentration.

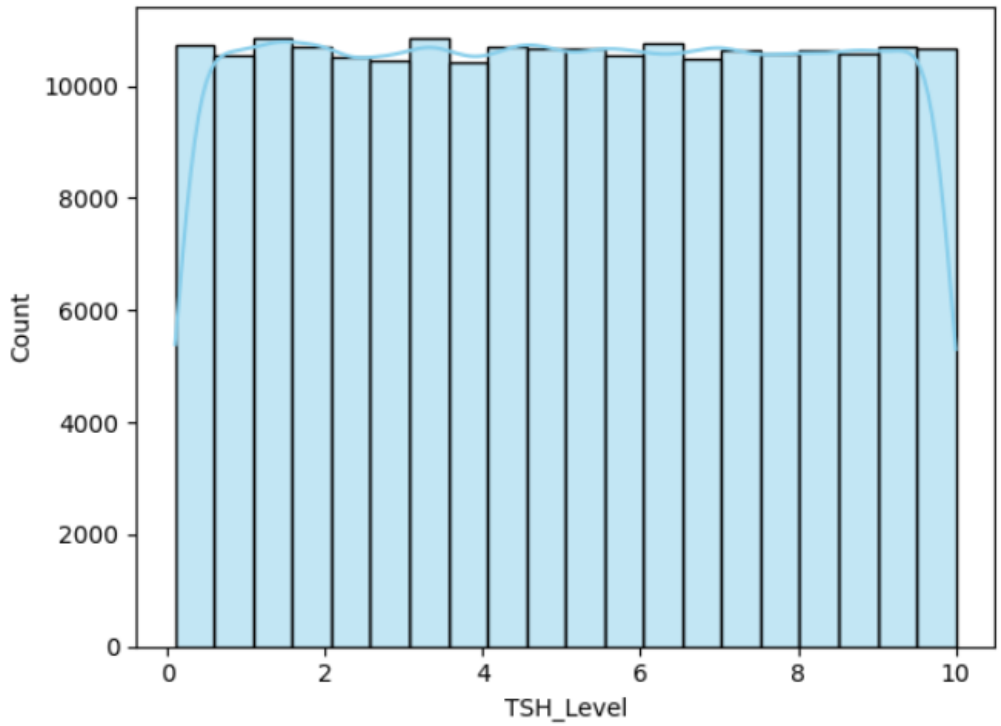
Age:

The Age distribution is fairly uniform across all ranges, indicating a balanced representation of younger, middle-aged, and older patients without significant skewness.

TSH_Level:

```
sns.histplot(dataset["TSH_Level"],kde=True, bins=20, color='skyblue')
```

```
<Axes: xlabel='TSH_Level', ylabel='Count'>
```

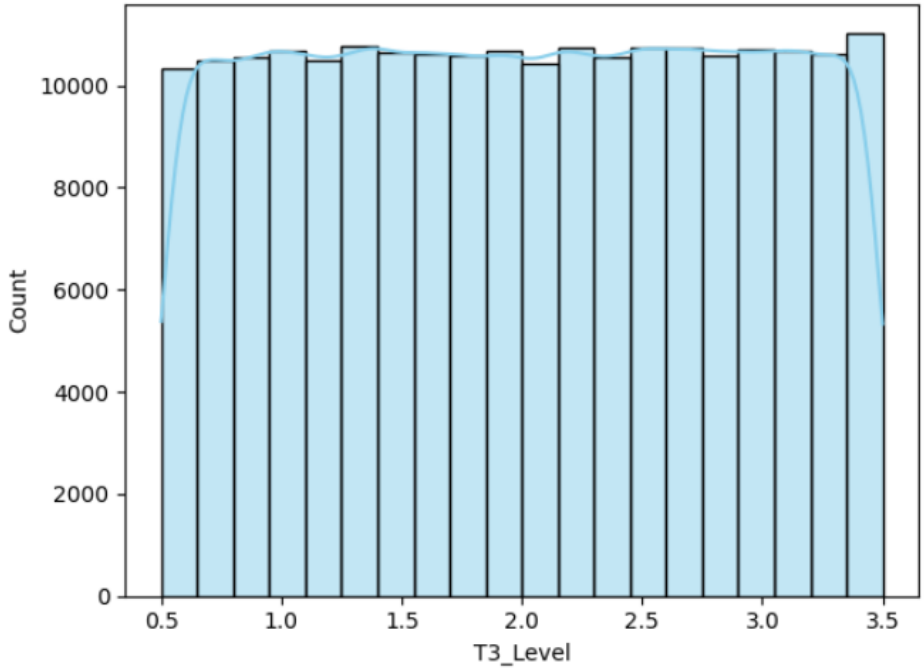


The distribution of **TSH_Level** appears fairly uniform with slight variations across the range, indicating that patients have a wide spread of TSH values. Most patients have normal to moderately elevated TSH levels, showing no strong skewness in the data.

T3_Level:

```
sns.histplot(dataset["T3_Level"],kde=True, bins=20, color='skyblue')
```

```
<Axes: xlabel='T3_Level', ylabel='Count'>
```

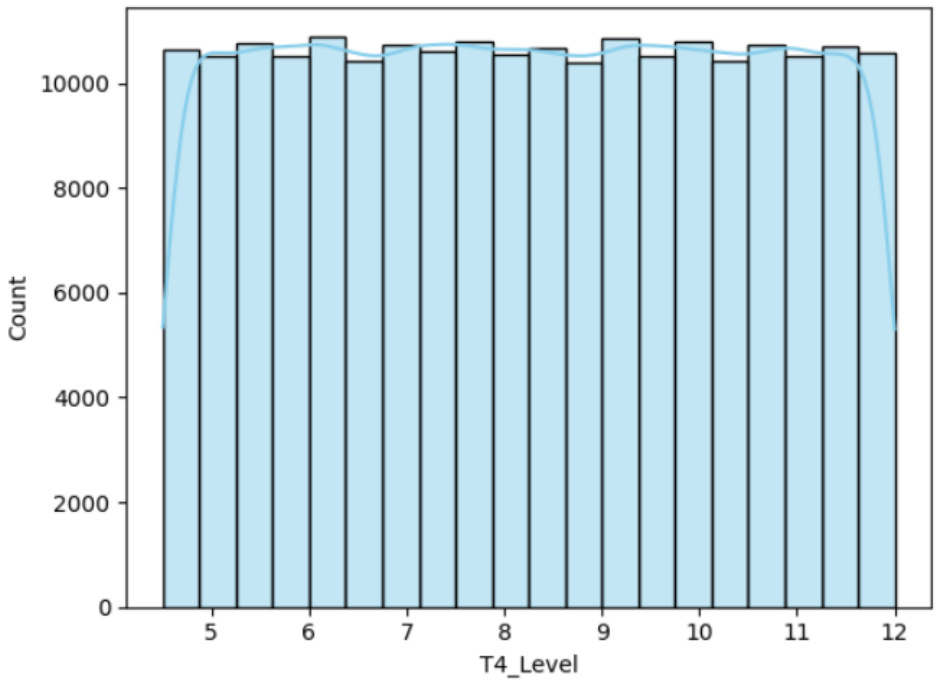


The distribution of **T3_Level** appears relatively uniform, suggesting that T3 values are evenly spread across patients. The data shows no strong skewness.

T4_Level:

```
sns.histplot(dataset["T4_Level"],kde=True, bins=20, color='skyblue')
```

```
<Axes: xlabel='T4_Level', ylabel='Count'>
```

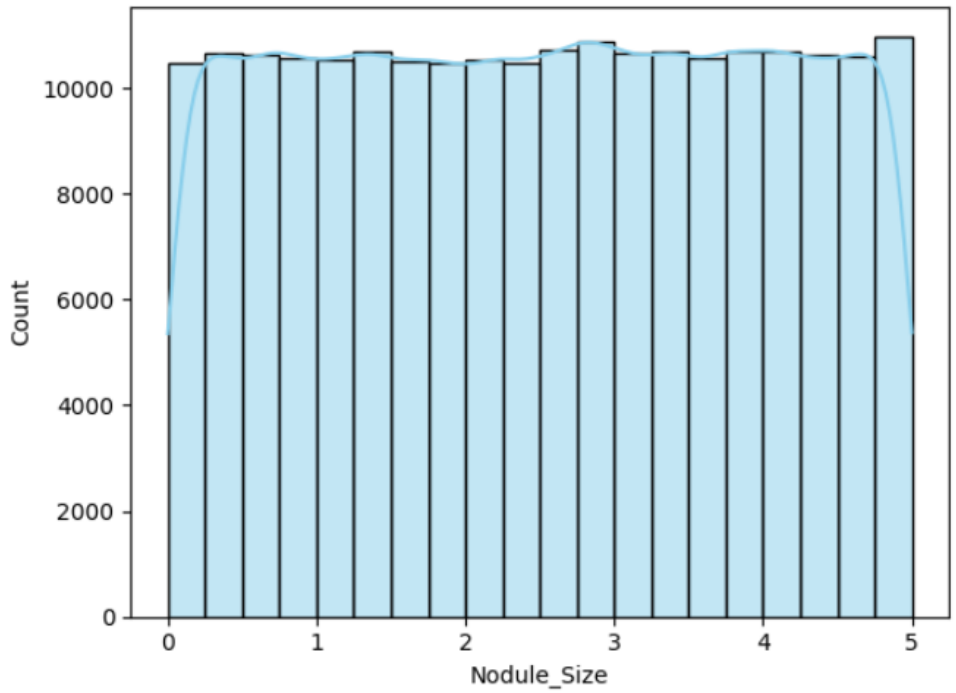


The distribution of **T4_Level** is fairly uniform across patients, indicating that T4 values are spread evenly without significant skewness. This suggests that most patients have T4 levels within the normal range.

Nodule size:

```
sns.histplot(dataset["Nodule_Size"],kde=True, bins=20, color='skyblue')
```

```
<Axes: xlabel='Nodule_Size', ylabel='Count'>
```



The distribution of **Nodule_Size** appears fairly uniform, indicating that nodule sizes are evenly spread across patients. This suggests there is no strong concentration of very small or very large nodules, and the dataset represents a balanced range of nodule measurements.

3. IQR (Inter Quartile Range)

The **IQR** helps identify **outliers** and understand the **spread** of the numerical data.

$$IQR = Q3 - Q1$$

	Age	TSH_Level	T3_Level	T4_Level	Nodule_Size
IQR	38	4.95	1.5	3.75	2.51
1.5rule	57	7.425	2.25	5.625	3.765
Lesser	-24	-4.855	-1	0.745	-2.515
Greater	128	14.945	5	15.745	7.525
Min	15	0.1	0.5	4.5	0
Max	89	10	3.5	12	5

By using formula in the python code we have found the lesser bound, Greater bound, Minimum value and Maximum value.

Age: Range: -24 to 128→ All ages (15 to 89) fall inside → So No Outliers

Age : IQR = 38 Min = 15 > Lesser = -24 Max = 89< Greater = 128 Conclusion: No outliers in Age
TSH_Level : IQR = 4.95 Min = 0.1 > Lesser = -4.8 Max = 10 < Greater = 14.9 Conclusion: No outliers in TSH_Level
T3_Level : IQR = 1.5 Min = 0.5 > Lesser = -1 Max = 3.5 < Greater = 5 Conclusion: No outliers in T3_Level
T4_Level : IQR = 3.75 Min = 4.5 > Lesser = 0.74 Max = 12 < Greater = 15.74 Conclusion: No outliers in T4_Level
Nodule_Size : IQR = 2.51 Min = 0 > Lesser = -2.5 Max = 5 < Greater = 7.5 Conclusion: No outliers in Nodule_Size

4. **Skewness and Kurtosis:**

Skewness: Measures the **asymmetry** of the data distribution.

Skew = 0 → Perfectly symmetrical

Skew < 0.5 → Fairly symmetric.

0.5 ≤ Skew< 1→ Moderately skewed.

Skew ≥ 1 →Highly skewed.

	Age	TSH_Level	T3_Level	T4_Level	Nodule_Size
Skew	0.004603	0.001629	-0.003318	0.001979	-0.004937
Kurtosis	-1.200477	-1.20093	-1.200503	-1.199327	-1.200245

The skewness values are close to zero, indicating that the data is nearly symmetric with no right or left tail.

Kurtosis:

It measures the **"tailedness"** of the distribution -> how heavy or light the tails are compared to a normal distribution.

Kurtosis = 0: Mesokurtic (normal distribution).

Kurtosis > 0: Leptokurtic — heavy tails, more outliers.

Kurtosis < 0: Platykurtic — light tails, fewer outliers

All kurtosis values are around **–1.2**, meaning the distributions are **platykurtic.**

Bivariate Analysis

1. **Correlation**:

Relation between two variables

Range: −1 to +1

- **+1 → Perfect positive correlation** (both increase together)
- **−1 → Perfect negative correlation** (one increases, the other decreases)
- **0 → No correlation** (no linear relationship)

	Age	TSH_Level	T3_Level	T4_Level	Nodule_Size
Age	1.000000	-0.000925	-0.001013	-0.002373	-0.001489
TSH_Level	-0.000925	1.000000	0.000335	-0.000795	0.000416
T3_Level	-0.001013	0.000335	1.000000	-0.004069	-0.001799
T4_Level	-0.002373	-0.000795	-0.004069	1.000000	-0.001860
Nodule_Size	-0.001489	0.000416	-0.001799	-0.001860	1.000000

The correlation values between all numeric features are **very close to 0**, which means there is **no significant linear relationship** between them.

2. Hypothesis Testing

Hypothesis testing checks whether there's **a significant relationship or difference** between two or more variables

- **Null Hypothesis (H_0):** There is *no significant difference or relationship.*
- **Alternative Hypothesis (H_1):** There *is* a significant difference or relationship.

ANOVA:

ANOVA (Analysis of Variance) is a **statistical test** used to check whether there are **significant differences between the means** of **three or more groups**.

```
# Here we are using oneway ANAVO
import scipy.stats as stats

# Split the data into groups
benign = dataset[dataset['Diagnosis'] == 'Benign']['T4_Level']
malignant = dataset[dataset['Diagnosis'] == 'Malignant']['T4_Level']

f_stat, p_value = stats.f_oneway(benign, malignant)

print("F-statistic:",f_stat)
print("P-value:",p_value)
```

```
F-statistic: 0.31424652168996675
P-value: 0.5750865561425021
```

- Conclusion

  if pvalue is < 0.05 we reject Null hypothesis

  if pvalue is > 0.05 we accept Null hypothesis (fail to reject)

  Here pvalue - 0.57 which is > than 0.05 ---> we are accepting Null hypothesis

  hence there is no significant difference between T4_Level and Diagnosis of patients.

T-test :

A **t-test** is a statistical test used to compare the **means (averages)** of two groups.

Type	**When to Use**
Independent (Unpaired) t-test	Compare means of **two different** groups (T4 level and diagnosis)
Paired t-test	Compare two sets of **related** measurements (before vs. after TSH levels)

```
from scipy.stats import ttest_ind

benign_pat= dataset[dataset['Diagnosis']=='Benign']['TSH_Level']
malignant_pat= dataset[dataset['Diagnosis']=='Malignant']['TSH_Level']

ttest_ind(benign_pat, malignant_pat)
```

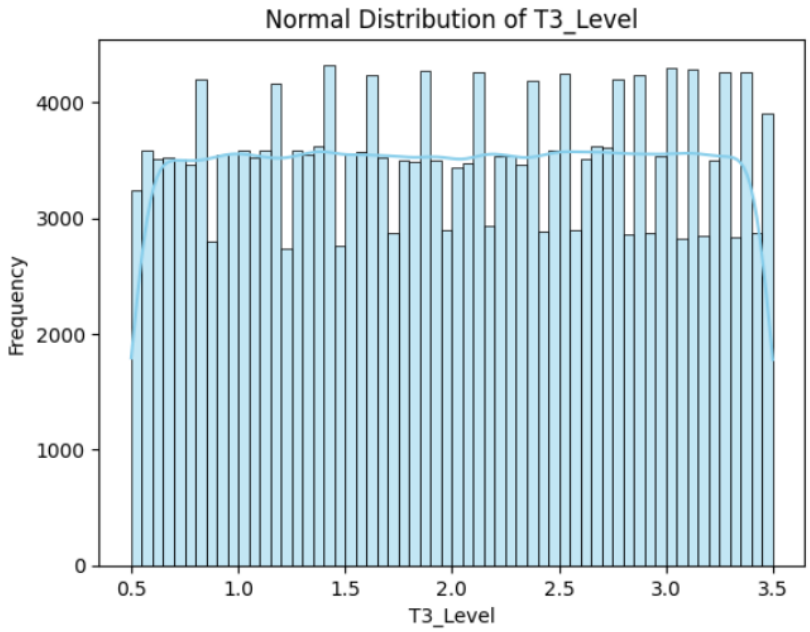
```
TtestResult(statistic=1.2312780727577617, pvalue=0.21822024145469007, df=212689.0)
```

```
#if pvalue is < 0.05 we reject Null hypothesis
#if pvalue is > 0.05 we accept Null hypothesis
#Here pvalue - 0.218 which is > than 0.05 ---> Hence we fail to reject Null hypothesis

#hence there is no significant difference between Diagnosis(Benign) and Diagnosis(Malignant) patients with respect to TSH_Level
```

Converting Normal distribution to Std Normal distribution:

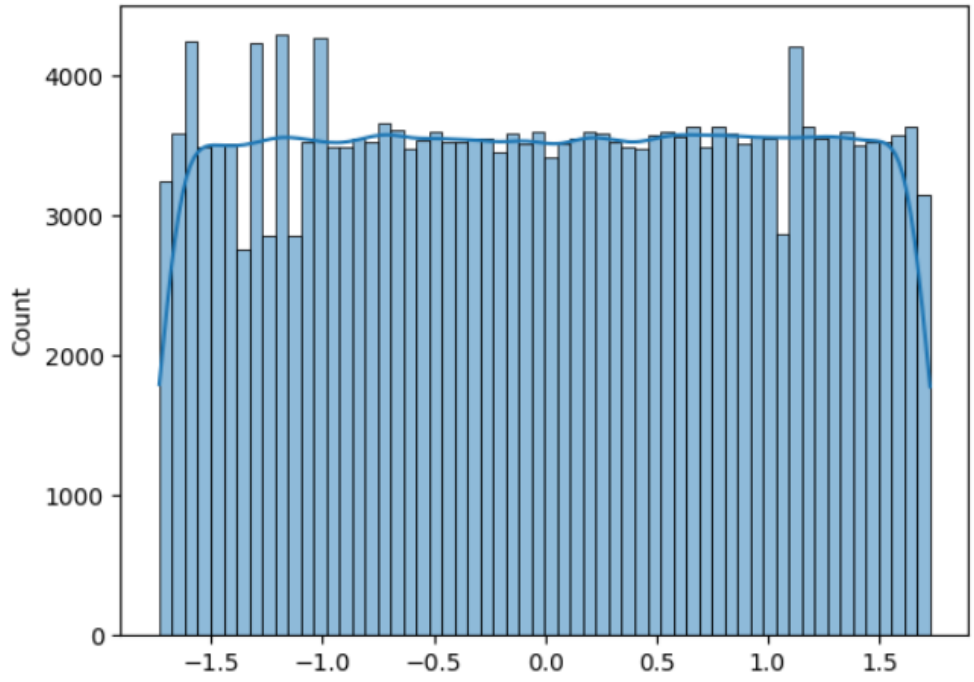
It is about **standardizing the data** — i.e., converting column (like TSH_Level, T3_Level, T4_Level,Age, etc.) to a **standard scale**.

```
import seaborn as sns
import matplotlib.pyplot as plt
sns.histplot(dataset["T3_Level"],kde=True, color='skyblue')

# Displaying the Normal Distribution of T3_Level
plt.title("Normal Distribution of T3_Level")
plt.xlabel("T3_Level")
plt.ylabel("Frequency")
plt.show()
```



```
# Coverted to standard Normal Distribution

mean=dataset["T3_Level"].mean()
std=dataset["T3_Level"].std()
values=[i for i in dataset["T3_Level"]]
z_score=[((j-mean)/std) for j in values]
sns.histplot(z_score,kde=True)
sum(z_score)/len(z_score)
```

```
-2.0671379627151948e-16
```



Data Visualization

EDA Questions and Answers

1. How many patients are diagnosed as Benign and Malignant?
    a. Benign 163196
    b. Malignant 49495

2. What is the percentage of patients by Thyroid_Cancer_Risk type?
    a. Low 50.960313
    b. Medium 34.039992
    c. High 14.999694

3. Display Top 10 patients with the highest Nodule Size?

```
dataset.sort_values(by='Nodule_Size',ascending=False).head(10)
```

	Age	Gender	Country	Family_History	Radiation_Exposure	Iodine_Deficiency	Smoking	Obesity	Diabetes	TSH_Level	T3_Level	T4_Level	Nodule_Size	Thyro
24061	18	Male	USA	No	No	No	No	No	No	6.73	1.77	9.61	5.0	
98394	81	Female	Brazil	No	No	No	No	Yes	Yes	5.46	2.24	7.41	5.0	
20947	56	Female	China	Yes	No	No	No	Yes	No	3.04	1.56	8.21	5.0	
117145	34	Female	India	No	No	No	No	Yes	No	1.15	0.51	9.60	5.0	
47803	44	Female	China	No	No	No	No	No	No	4.35	3.07	5.71	5.0	
57305	18	Male	Nigeria	Yes	No	Yes	Yes	No	No	3.22	3.50	5.10	5.0	
172015	60	Male	Russia	No	No	Yes	No	No	Yes	1.21	1.59	8.14	5.0	
75226	88	Male	Brazil	No	No	Yes	No	Yes	No	3.83	0.65	8.91	5.0	
183242	48	Female	South Korea	No	Yes	Yes	No	No	No	0.89	2.40	11.82	5.0	
110153	33	Male	Russia	No	No	Yes	No	No	No	4.88	0.81	7.55	5.0	

4. Which patient from India has the largest Nodule Size?

```
india_patients = dataset[dataset['Country'] == 'India']
india_patients.sort_values(by='Nodule_Size',ascending=False).head(1)
```

	Age	Gender	Country	Family_History	Radiation_Exposure	Iodine_Deficiency	Smoking	Obesity	Diabetes	TSH_Level	T3_Level	T4_Level	Nodule_Size	Thyroid
68074	54	Male	India	No	No	No	No	Yes	No	5.52	2.39	8.79	5.0	

5. Which countries have the highest *TSH-Level* for each *Diagnosis* type (Benign/ Malignant)?

Diagnosis	Country	TSH_Level
Benign	Brazil	5.075435
Malignant	Japan	5.150928

6. What is the average nodule size distribution of the high-Thyroid_Cancer_Risk for each Diagnosis type?

Diagnosis	Nodule_Size
Benign	2.506930
Malignant	2.498084

7. Which countries have the highest number of cases for each Diagnosis type?

```
top_countries = case_counts.groupby('Diagnosis').head(2)
print(top_countries)
```

	Diagnosis	Country	Case_Count
3	Benign	India	28520
1	Benign	China	25280
13	Malignant	India	13976
15	Malignant	Nigeria	6712

8. Is there a relationship between Smoking and Thyroid Cancer Risk?

```
# Count how many patients per Smoking-Risk group
grouped = (
    dataset.groupby(['Smoking', 'Thyroid_Cancer_Risk'])
    .size()
    .reset_index(name='Count')
)
print(grouped.head())
```

```
  Smoking Thyroid_Cancer_Risk  Count
0      No                High  25534
1      No                 Low  86658
2      No              Medium  58068
3     Yes                High   6369
4     Yes                 Low  21730
```

9. Is there a relationship between Obesity and Thyroid Cancer Risk?

```
  Obesity Thyroid_Cancer_Risk  Count
0      No                High  22430
1      No                 Low  75766
2      No              Medium  50609
3     Yes                High   9473
4     Yes                 Low  32622
```

10. Is there a relationship between Iodine deficiency and Thyroid cancer risk?

```
Iodine_def = (
    dataset.groupby(['Iodine_Deficiency', 'Thyroid_Cancer_Risk'])
    .size()
    .reset_index(name='Count')
)
print(Iodine_def.head())
```

```
  Iodine_Deficiency Thyroid_Cancer_Risk  Count
0                No                High  16846
1                No                 Low  85630
2                No              Medium  57197
3               Yes                High  15057
4               Yes                 Low  22758
```