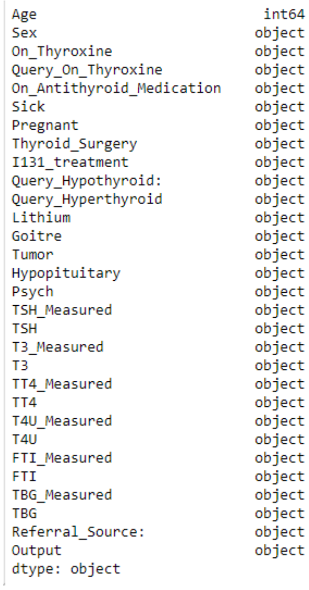
Thyroid Detection EDA and Preprocessing

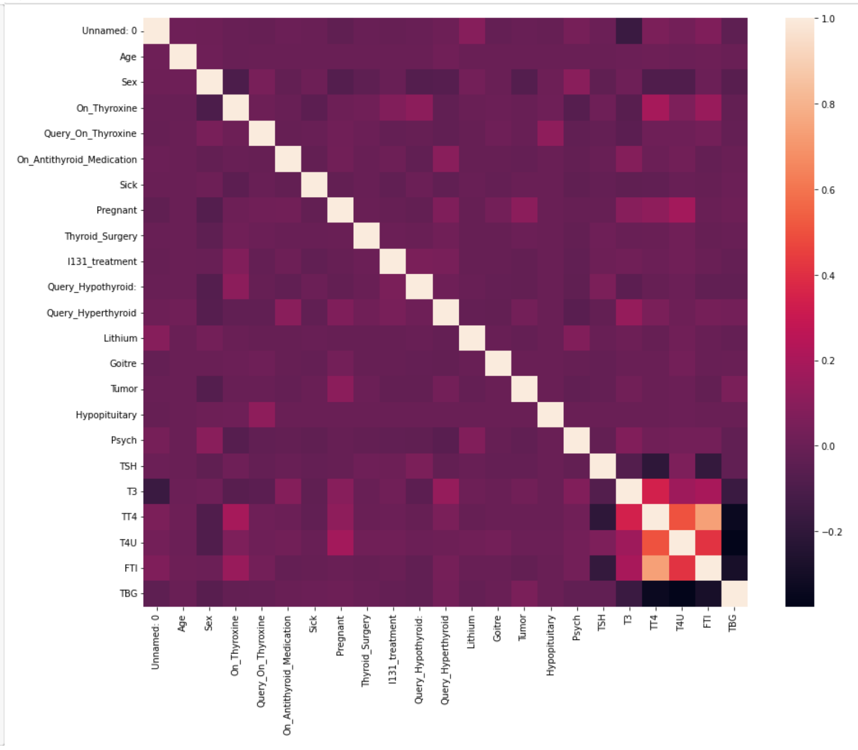
1. Data Provided Contains the 9172 rows and 30 Columns
2. The 30 Columns are as below:
    a. 'Age','Sex','On_Thyroxine','Query_On_Thyroxine','On_Antithyroid_Medication','Sick', 'Pregnant','Thyroid_Surgery','I131_treatment','Query_Hypothyroid:','Query_Hyperthyroid','Lithium','Goitre','Tumor','Hypopituitary','Psych','TSH_Measured','TSH','T3_Measured','T3','TT4_Measured','TT4','T4U_Measured','T4U','FTI_Measured','FTI','TBG_Measured','TBG','Referral_Source:','Output'
    b. Where output is dependent columns are other columns are independent columns.
3. Datatype of the columns are as follows:



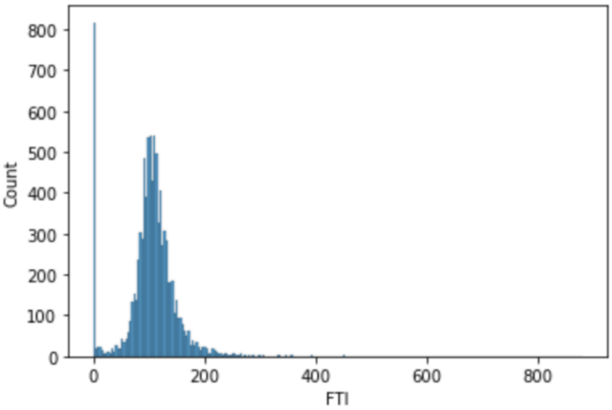
```
Age                           int64
Sex                          object
On_Thyroxine                 object
Query_On_Thyroxine           object
On_Antithyroid_Medication    object
Sick                         object
Pregnant                     object
Thyroid_Surgery              object
I131_treatment               object
Query_Hypothyroid:           object
Query_Hyperthyroid           object
Lithium                      object
Goitre                       object
Tumor                        object
Hypopituitary                object
Psych                        object
TSH_Measured                 object
TSH                          object
T3_Measured                  object
T3                           object
TT4_Measured                 object
TT4                          object
T4U_Measured                 object
T4U                          object
FTI_Measured                 object
FTI                          object
TBG_Measured                 object
TBG                          object
Referral_Source:             object
Output                       object
dtype: object
```

4. Age values ranges from 1 to 65526
5. Sex will contains the values either Male(M), Female(F), or ?(pd.nan)
6. ‘On_Thyroxine’ will have the value either ‘T’ or ‘F’
7. ‘Query_On_Thyroxine’ will have the value either ‘T’ or ‘F’
8. ‘On_Antithyroid_Medication’ will have the value either ‘T’ or ‘F’
9. ‘Sick’ will have the value either ‘T’ or ‘F’
10. ‘Pregnant’ will have the value either ‘T’ or ‘F’
11. ‘Thyroid_Surgery’ will have the value either ‘T’ or ‘F’
12. ‘I131_treatment’ will have the value either ‘T’ or ‘F’
13. ‘Query_Hypothyroid’ will have the value either ‘T’ or ‘F’
14. ‘Query_Hyperthyroid’ will have the value either ‘T’ or ‘F’
15. ‘Lithium’ will have the value either ‘T’ or ‘F’
16. ‘Goitre’ will have the value either ‘T’ or ‘F’
17. ‘Tumor’ will have the value either ‘T’ or ‘F’
18. ‘Hypopituitary’ will have the value either ‘T’ or ‘F’
19. ‘Psych’ will have the value either ‘T’ or ‘F’
20. ‘TSH_Measured’ will have the value either ‘T’ or ‘F’
21. ‘T3_Measured’ will have the value either ‘T’ or ‘F’
22. ‘TT4_Measured’ will have the value either ‘T’ or ‘F’
23. ‘T4U_Measured’ will have the value either ‘T’ or ‘F’
24. ‘FTI_Measured’ will have the value either ‘T’ or ‘F’

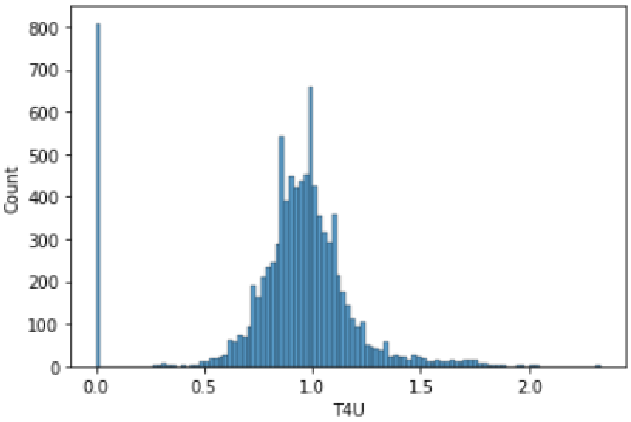
25. ‘TBG_Measured’ will have the value either ‘T’ or ‘F’
26. ‘Referral_Source’ have the values 'other', 'SVI', 'SVHC', 'STMW', 'SVHD', 'WEST'
27. Dropping ‘Referral_Source’ as the Data present in the column is not valid.
28. In the Data set we have ? as the unknow values hence replacing them with pd.nan values
29. Columns contains the null data in the following order:
    a. Sex has 307 null values
    b. TSH has 842 null values
    c. T3 has 2603 null values
    d. TT4 has 442 null values
    e. T4U has 809 null values
    f. FTI has 802 null values
    g. TBG has 8823 null values
30. ‘TBG’ has 8823 null values hence decided the to drop the column
31. ‘TBG_Measured’ will also be dropped as we have dropped ‘TBG’
32. Values marked with ‘f’ and ‘t’ will be replaces with ‘0’ and ‘1’ respectably.
33. The Null values present in the columns ‘TSH’, ‘T3’,’TT4’,’T4U’,’FTI’ can be marked as 0 if the corresponding value with measured column marked as ‘f’
34. Dropping all the columns with Measured.
35. After the above data wrangling, we are only ‘Sex’ Column is left with 307 null values
36. Creating the Dummies for the column ‘Sex’ where null values will be marked as np.nan, ‘M’ as 1 and ‘F’ as 0.
37. Use ‘KNN imputer’ to find the missing values for the column ‘Sex’.
38. Checking the correlation of the Columns:



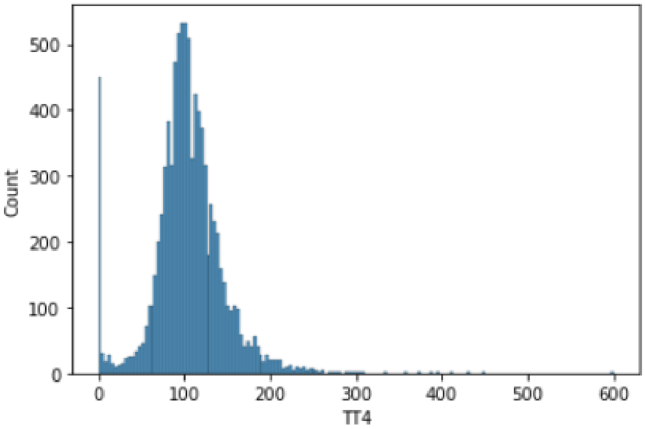
39. Check the distribution of the Measured Columns:
    a. FTI is normally distributed:



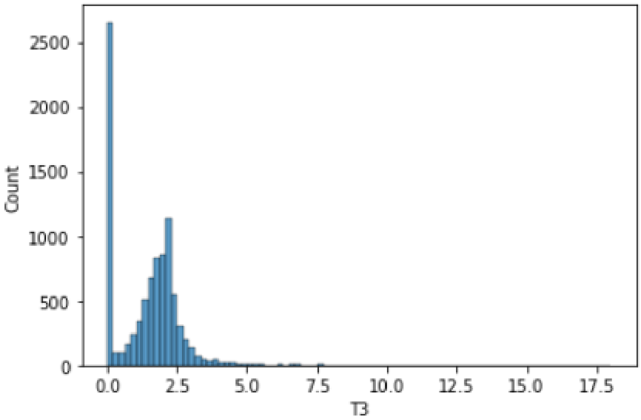
b. T4U is normally distribute :



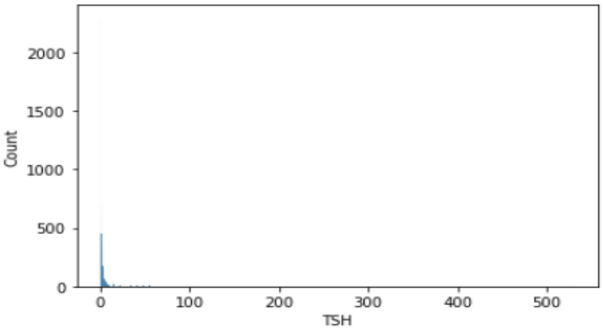
c. TT4:



d. T3:

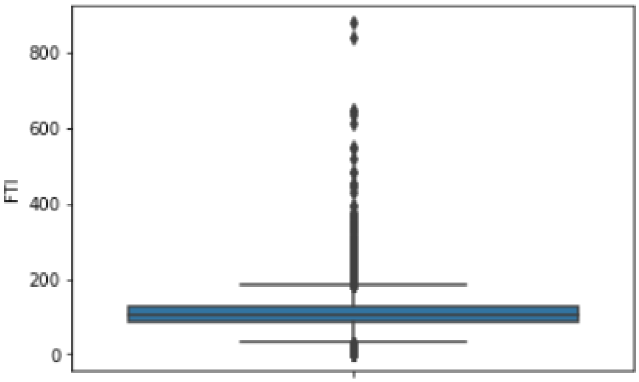


e. TSH data is right skewed :

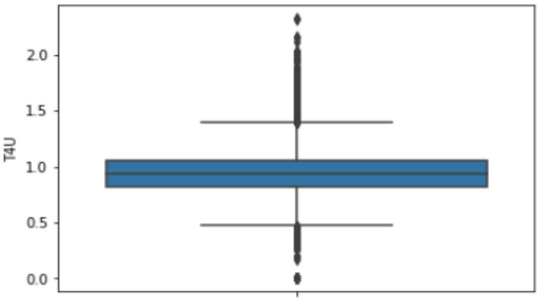


40. Checking for the outlier of the measured Columns:

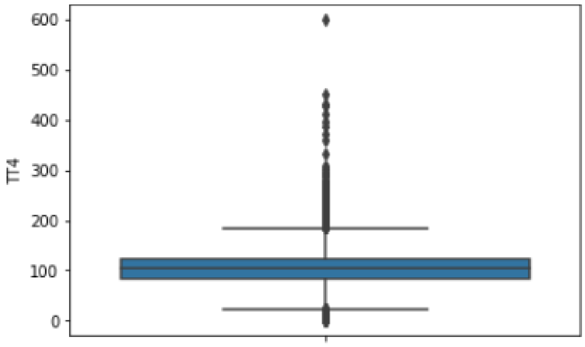
    a. FTI



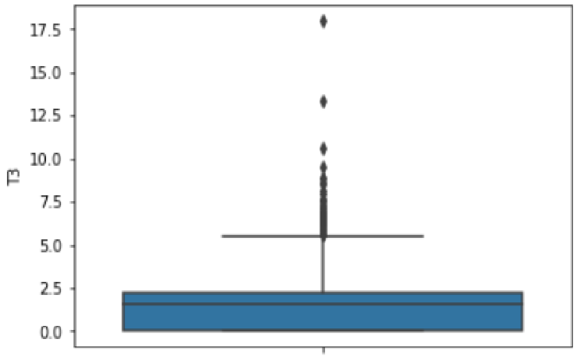
    b. T4U



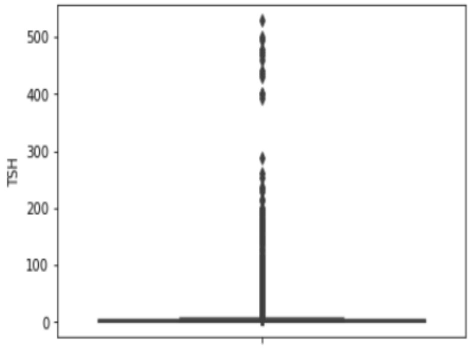
c. TT4



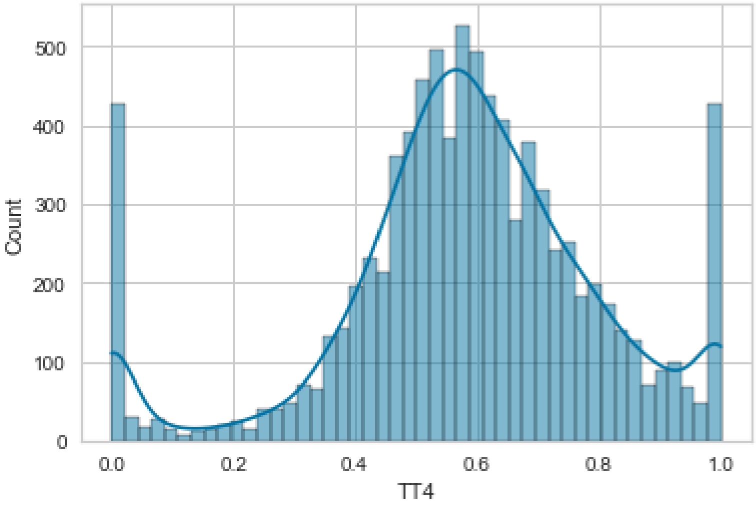
d. T3:



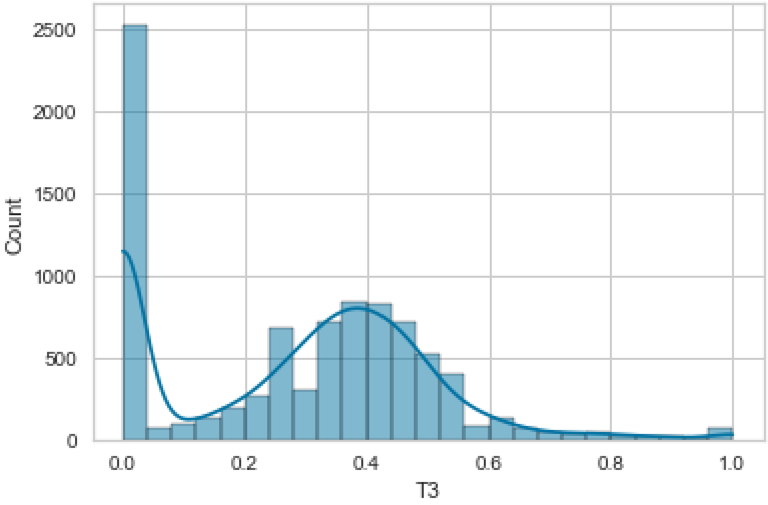
e. TSH:



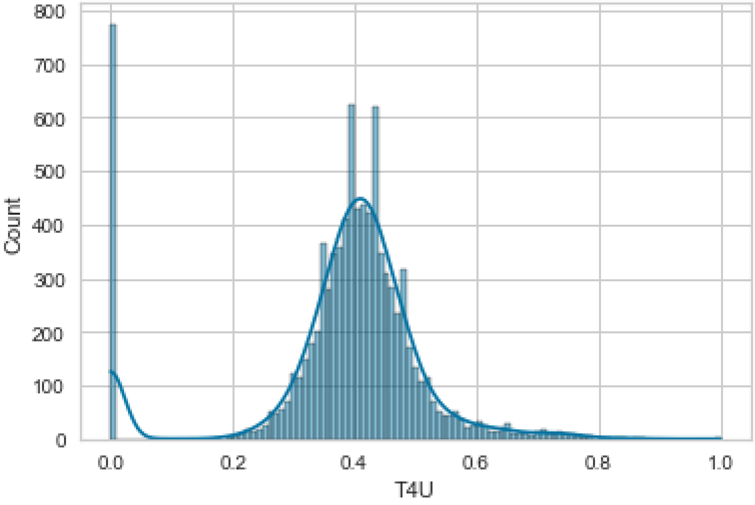
41. After checking the above data shown in the point 40 below points are considered:
    a. ‘TBG’ values above 50 will be considered as 50
    b. ‘TSH’ values above 5 will be considered as 5
    c. ‘T3’ values above 5 will be considered as 5
    d. ‘TT4’ above 174 will be considered as 175
    e. ‘FTI’ above 200 will be considered as 200
42. Since the data varies between the large scale we need to skew the data between the range of 0 and 1.
    a. TT4:



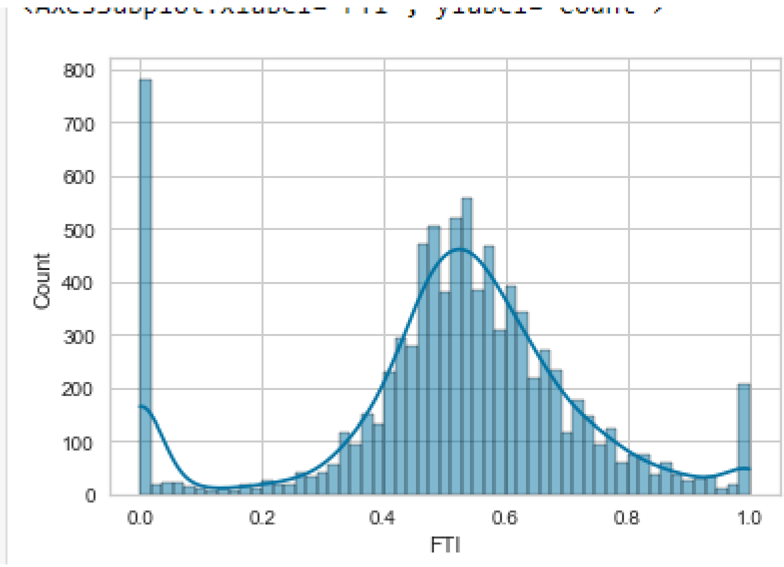
b. T3:



c. T4U:

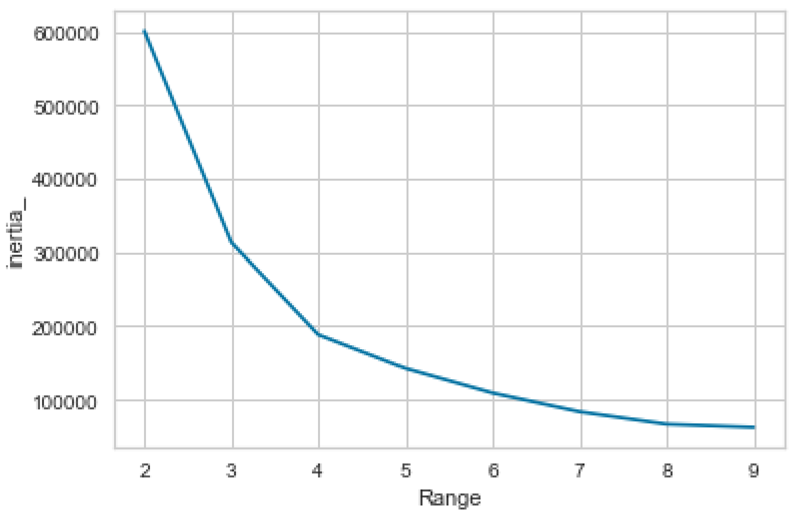


d. FTI:

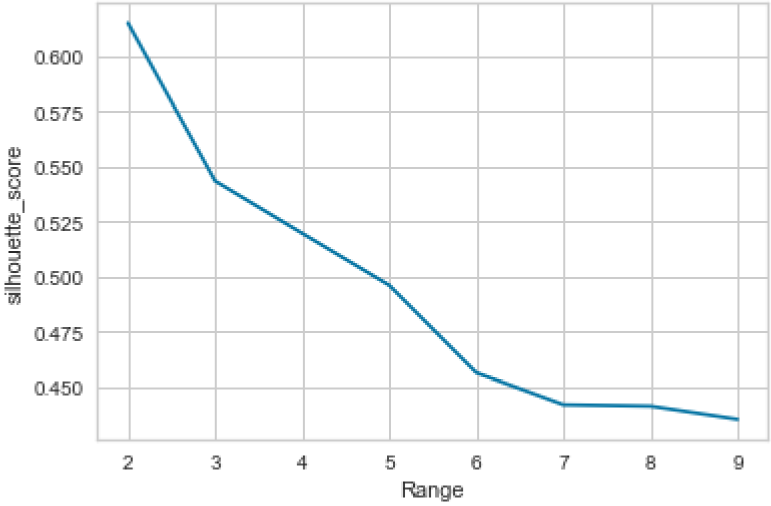


```
<AxesSubplot:xlabel='FTI', ylabel='Count'>
```

43. The above shown data are in the normally distributed manner.
44. We will now use the data to make the clusters of the data for which I will be using the Kmeans Clustering.

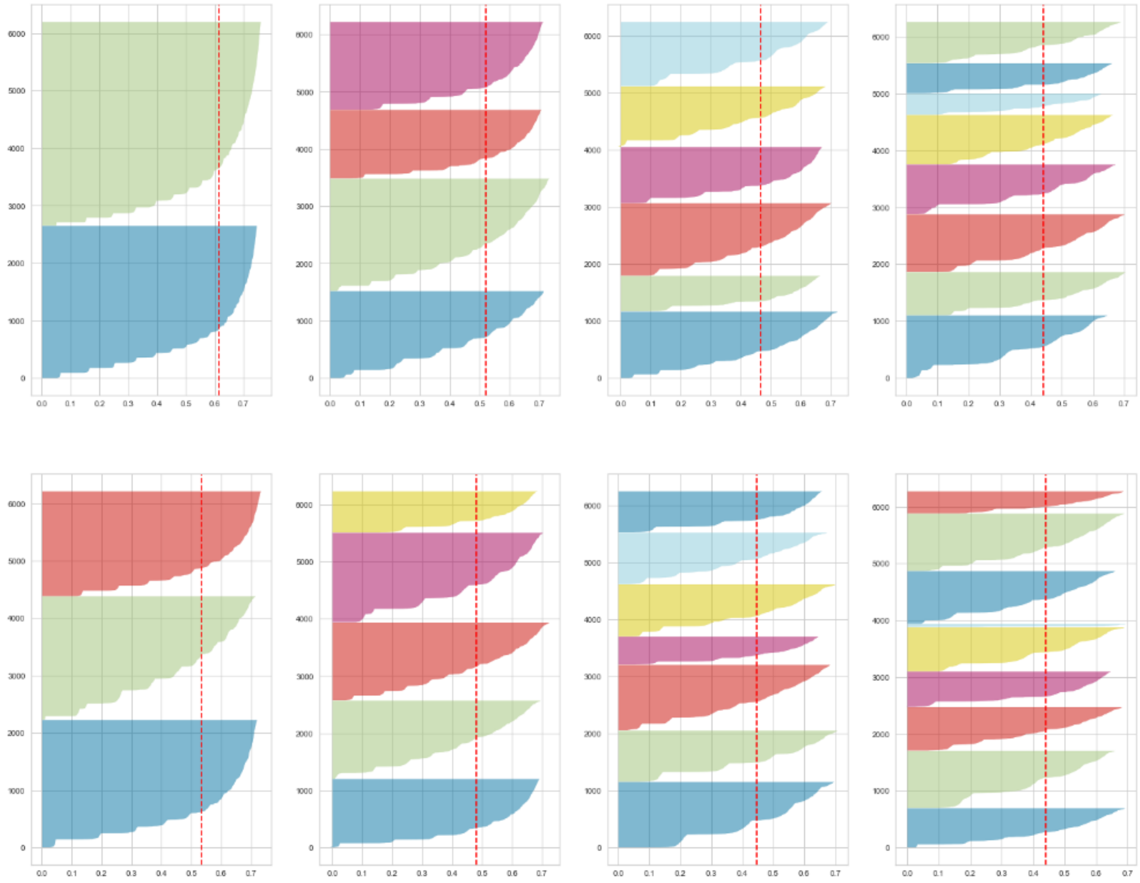


a.
b. The above graph represents the inertia and clusters which can be verified by silhouette_score



c.

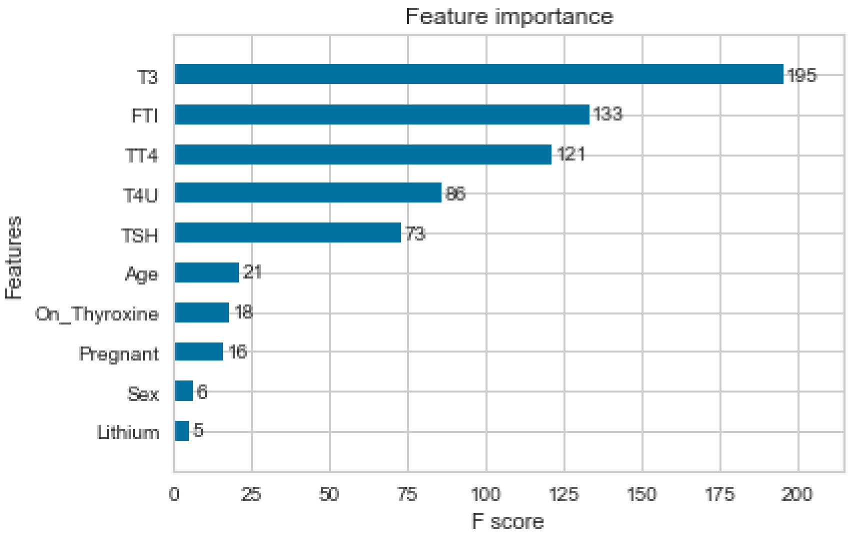
d. Drawing the clusters :



Optimized values of the clusters will be 4 as the values in the above diagram is greater then 0.5.

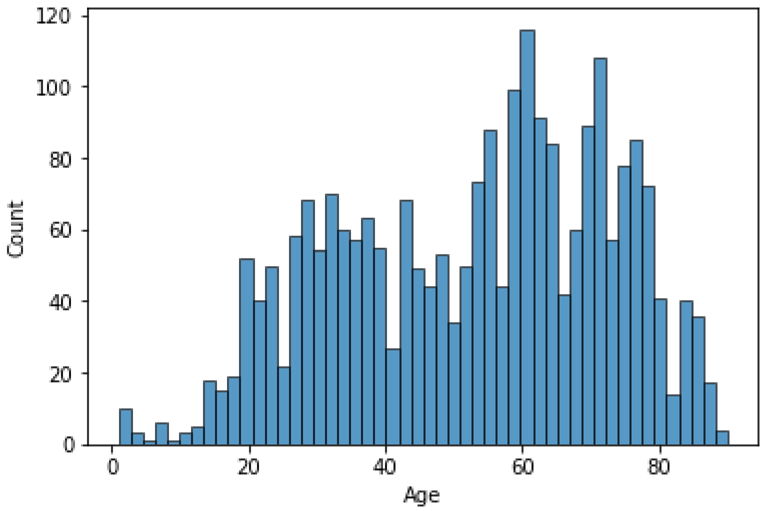
45. Top 10 features that are impacting the output variables are as follows(using XGBoosting):

```
<AxesSubplot:title={'center':'Feature importance'}, xlabe
```



a.

46. Age distribution of the people having Thyroid:



People with Age group between 20 to 80 are having the chances of thyroid more.