# Springboard data science intensive capstone project

# Thyroid classification.

Rahul Ambati

1. Introduction.
2. Data acquisition and wrangling.
3. Data exploration and inferential statistics.
4. Summary and further steps.

# Introduction.

*Thyroid disease* is a medical condition affecting the function of the thyroid gland. And the symptoms of the disease vary depending on the type of thyroid disease.

Intention:

A physician needs to know the demographics suffering from thyroid disease and find what sector of people can be focused on so that they get admitted and get prior treatment.

Client:

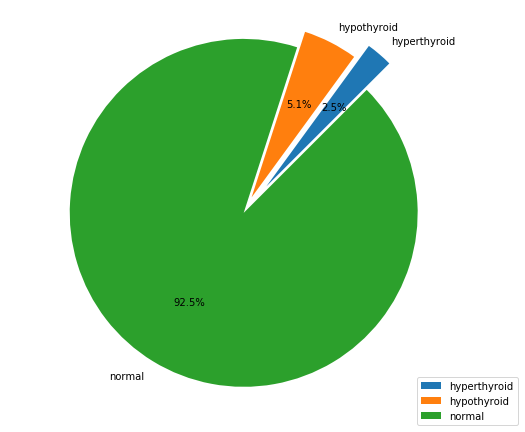
*Physicians* who want to understand what kind of demographics, medication, etc. to consider while treating patients with thyroid disease so that the right group of people can get the proper care and treatment.

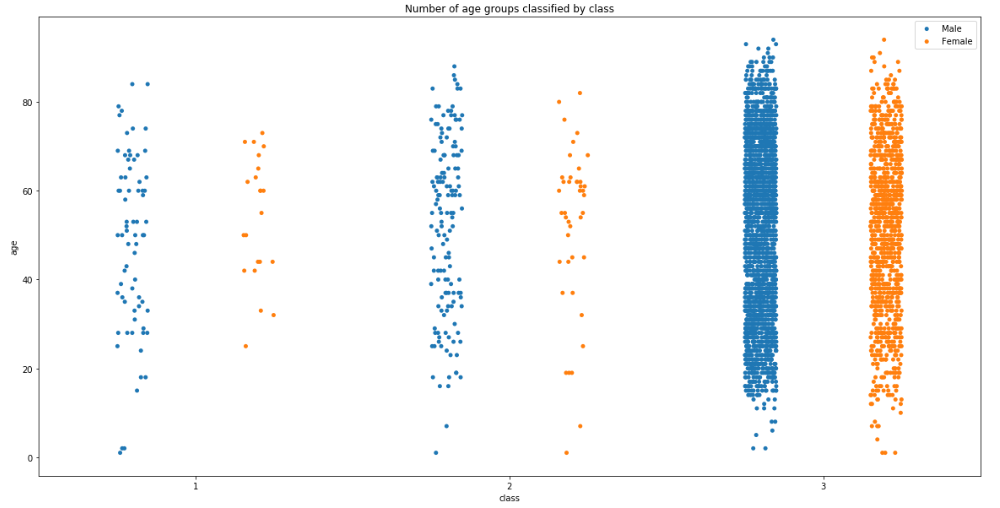
# Data acquisition and wrangling.

The data is acquired from [UCI ML dataset Thyroid disease](https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease) and in particular will be focusing on [ANN](https://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/ann-Readme). ANN dataset provides the information for finding thyroid disease.  
  
Dataset consists of demographics (age, sex), queries about medication, current conditions (sick, tumor, goiter, etc.) and measurements (TSH, T3, TT4, T4U, FTI) and category.

The dataset from the repository is clean, not much cleaning or wrangling had to be done.

# Data exploration and inferential statistics.

There are 3 categories ***hyperthyroid*** *(class 1),* ***hypothyroid*** *(class 2) and* ***normal*** *(class 3)*. ***Normal*** category means person does not suffer from any thyroid disease.  
  
Exploring the data, much of the demographics does not suffer from any thyroid disease (Figure 1).  
  
Figure 1.

From the data set a strip plot (Figure 2) is made to view the category and age group.Figure 2.

Density plots (Figure 3) for TSH, T3, TT4, T4U, FTI measurements are made, and see where the distributions lie.

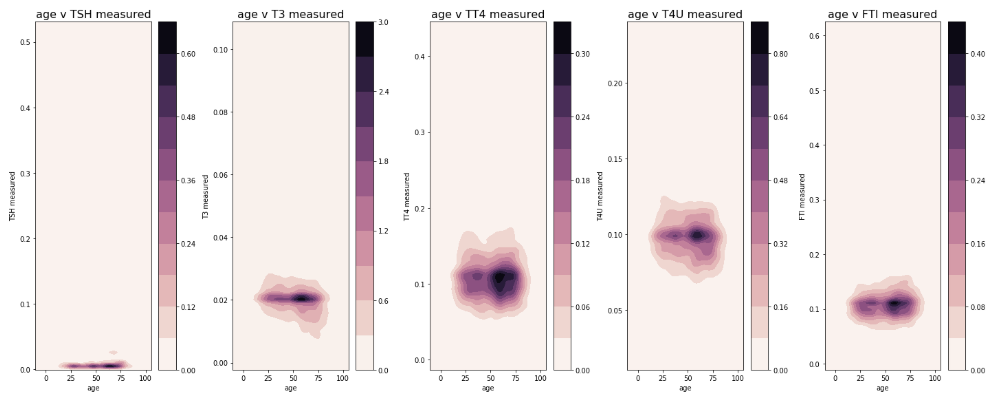
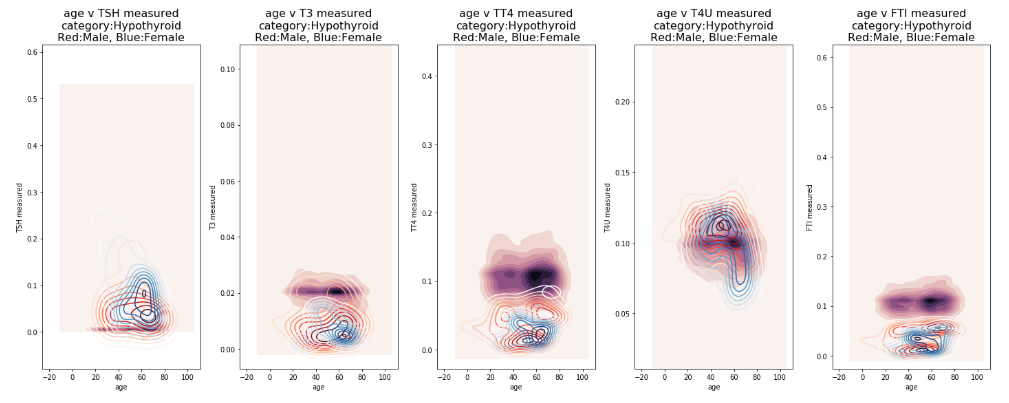


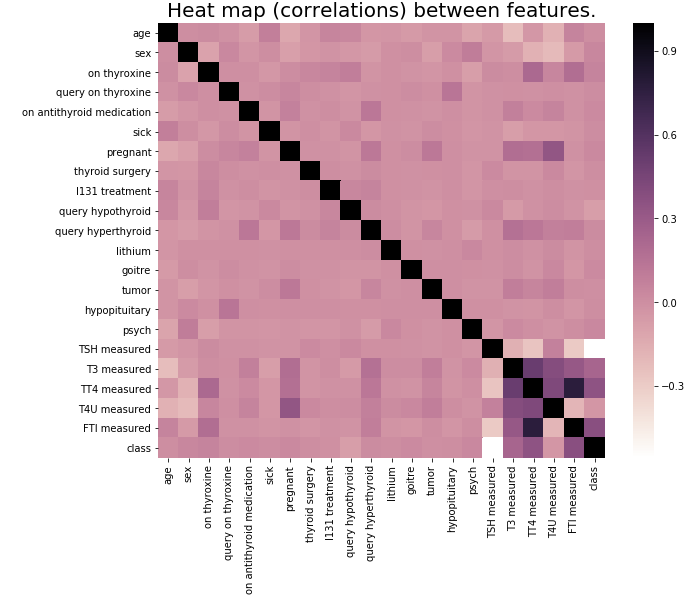
Figure 3.

***Hyperthyroid*** class was extracted from the data set and overlapped on the measurements to see where it lies in the distribution (Figure 4).

Figure 4.

A clear out liners can be seen where the ***hyperthyroid*** lies and we can make some inferences that it might lie outside the normal distributions.

With many and different features each feature can affect differently, so a correlation map is made to see how features are affecting the most and the least (Figure 5).

   
Figure 5.

Looking at the correlation map (Figure 5) the measurements have a varied level of influence.

***Statistical inference*** was used to determine if correlation between TT4 and FTI measurements are statistically significant, by performing t-test.

*Null hypothesis* mentions correlation between TT4 and FTI measurements is zero.

# Summary and further steps.

Data collection, wrangling is done. Exploratory data analysis is performed on the data set.

Next step(s) will be applying machine learning techniques in predicting classes.