ST494 Final Project
Thyroid Disease

April.5 2024

Yiqian Kang

Executive Summary

Our project is dedicated to improving Hypothyroidism prediction capabilities, aiming to accurately predict a patient's likelihood of developing hypothyroidism following medical intervention. The dataset we used is complex, consisting primarily of categorical variables, mostly using binary (true or false) indicators, and exhibits a few key numerical variables that provide a deeper understanding of each patient's health status. Given the limited size of the dataset (containing 3,163 observations), and the challenges of predicting hypothyroidism, our approach requires careful investigation and processing of the data. Additionally, a key decision in our project revolves around the minimization of errors, specifically false positives and false negatives. Recognizing the profound impact of each type of error, we strive to simultaneously minimize both by achieving models with the highest possible accuracy.

When we started this project, we split the dataset into a 70% training set and 30% test set, this split facilitated testing and validation of the model. Through the project, we aim not only to improve the prediction accuracy of hypothyroidism but also to determine the most important predictor variables. Our project will use comprehensive data analysis to better predict and understand hypothyroidism.

Problem

Our data set focuses on hypothyroidism, also known as underactive thyroid, where the thyroid gland does not produce enough thyroid hormone to meet the body's needs. Thyroid hormones control how your body uses energy, so they affect nearly every organ in your body, even the way your heart beats. Without enough thyroid hormone, many of the body's functions slow down and can lead to a range of health problems including slowed heart rate, weight gain, joint and muscle pain, fatigue, or fertility problems. The complexity of this disease stems from its multifaceted nature, which is influenced by factors such as age, gender, hormone levels and medication use. The data can be found at this site: http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease

Data Description

Categorical Variables (Bool)

- **Sex**: Indicates the patient's gender (Str: Male, Female).
- **On_thyroxine**: Whether the patient is currently taking thyroxine medication.

- **Query_on_thyroxine**: Indicates queries related to thyroxine medication usage.
- **On_antithyroid_medication**: Whether the patient is on antithyroid medication.
- **Thyroid_surgery**: Indicates if the patient has undergone thyroid surgery.
- **Query_hypothyroid**: Queries related to hypothyroid conditions.
- **Query_hyperthyroid**: Queries related to hyperthyroid conditions.
- **Pregnant**: Indicates if the patient is pregnant.
- **Sick**: Reflects if the patient was sick at the time of data collection.
- **Lithium**: Indicates if the patient has taken lithium.
- **Goitre**: Indicates if the patient has a goiter.
- **Tumour**: Indicates if the patient has a tumour.
- **TSH_measure**: Whether TSH was measured in blood.
- **T3_measure**: Whether T3 was measured in blood.
- **TT4_measure**: Whether TT4 was measured in blood.
- **T4U_measure**: Whether T4U was measured in blood.
- **FTI_measure**: Whether FTI was measured in blood.
- **TBG_measure**: Whether TBG was measured in blood.

Continuous Variables (Float)

- **Age**: The patient's age, a crucial factor as thyroid disorders can vary with age. (Int)
- **TSH** (Thyroid Stimulating Hormone): TSH level in blood from lab work.
- **T3** (Triiodothyronine): Measures the active thyroid hormone level.
- **TT4** (Total Thyroxine): Reflects the total amount of thyroxine in the blood.
- **T4U** (Thyroxine Uptake): Indicates the amount of thyroxine being taken up by cells.
- **FTI**(Free Thyroxine Index): Calculated from TT4 and T4U, providing an estimate of free thyroxine levels.
-**TBG** (Thyroxine-Binding Globulin): Although often not directly measured, it's crucial for interpreting total hormone levels.

Target Variable

- Diagnosis: The primary outcome variable, indicating whether the patient has been diagnosed with a hyperthyroid condition within a year after undergoing specific medical interventions or tests.

Project Objectives

• Our overarching goal with this project is to be able to accurately predict whether a patient will be diagnosed with hyperthyroidism within one year of receiving a specific medical intervention or test.

• Another goal is that through various models and methods we hope to be able to determine which factors are most important in determining the response variable in our data set.

• Also check which method will give us better performance.

Dataset Problem

We have 5,328 missing entries in our dataset, so we must use a reliable method to handle all missing values and retain all valuable information. We first try to remove columns with higher missing values. Initially, the data set has 26 columns. After removing columns with more than 50% missing data ("Threshold <- 0.5"), the dataset has 3,162 observations and 25 variables. Considering the impact of row-wise deletion on dataset size and information loss, we adopt an imputation strategy that retains more data. From the histograms and boxplots as well as the skewness values of each variable, we found that TSH, T3, TT4, T4U, and FTI, these thyroid-related measurements, showed right skewness. So we will use the median for interpolation since it is less affected by extreme values. The remaining variables (binary) are categorical and therefore do not require imputation.

Outlier Problem

In our dataset, a significant presence of outliers is observed across various parameters: TSH has 438 outliers, T3 with 257, TT4 with 252, T4U with 248, and FTI with 299. Given the substantial number of outliers and their potential clinical relevance to hyperthyroidism, we will not disregard them hastily. These outliers might signify severe instances of the disease or exceptional reactions to therapies. Therefore, for the integrity of our study and to preserve the clinical significance of the data, the outliers will remain unaltered at this juncture.

Train/Test Set Decision

With our dataset having roughly 3163 observations, we decided to not use a validation set but rather create a stratified test and train set. The training set included 70% of the data and the test set contained the remaining 30%. For some models we elected to use cross validation to make sure the results were accurate and robust.

PCA Original data (unscaled):

The first principal component (PC1) explained 40.82% of the variance, and the first two components (PC1 and PC2) combined explained 65.18%.

The loadings for PC1 show that TT4 and FTI have the largest absolute values, meaning they have the greatest impact on PC1. For PC2, Age and T4U have the greatest impact.

PCA Transformed and scaled data:

When PCA was performed on the transformed and scaled data, the first component explained 43% of the variance, followed by 24.82% of the second component, and so on. This is a more even distribution compared to the unscaled PCA, and we note that for PC1: the largest negative loading is TT4, followed by FTI and T3, indicating their influence in the negative direction of PC1. Age has a positive loading, but it is not the most important. PC2: Age shows a strong positive loading, while T4U has the strongest negative loading. This contrast may indicate that PC2 captures changes in the dataset associated with age-related changes that are distinct from thyroid function as represented by T4U.

After transformation and scaling, PCA presents a more balanced view, with different variables showing significant contributions from individual components. Given the lower cumulative variance captured by the initial components in scaled PCA, models using this data are likely to have poorer predictive power. So we will use the converted data for subsequent testing

Clustering Results

Our cluster plot shows four clusters that are reasonably well separated in the two-dimensional space formed by the first two principal components. This shows that distinct groups exist in the data and are clearly separated, proving that the clusters are well structured and distinguishable.

K-means suggests 3 or 4 clusters based on the elbow method, PCA clustering visually supports 4 clusters.

We first try to cluster the results on the original dataset. We try to use hierarchical clustering to find clustered data. We perform this clustering using single, complete, and average methods. All clusters failed to achieve good results. We then ran the same clustering method on the dataset after PCA transformation. The clustering results after PCA transformation are also poor, and the hierarchical clustering method cannot provide us with any useful information.

LDA and QDA Results

The analysis using Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) has produced the following results:

	Misclassification Rate	F1 score	Precision	Recall
LDA	2.4%	0.99	0.98	1
class 0	/	0.66	0.88	0.53
class 1	/	0.99	0.98	1
QDA	95.7%	0	0	0
class 0	/	0.08	0.04	1
class1	/	0	0	1

LDA provides high accuracy and performance metrics, indicating that it is a good model for this dataset. QDA performs poorly, which may be due to the data not meeting the underlying assumptions of QDA.

SVM Results

Through the SVM model, we obtained an accuracy of 0.895, which shows that our SVM model is highly reliable in identifying positive cases. Because it minimizes the number of false positives (ex: situations where patients are incorrectly diagnosed with a disease they don't have). The recall rate is 0.773, our recall rate is relatively high, indicating that the model successfully captures a large proportion of hypothyroidism cases, and the F1 score is 0.829 indicating a good balance between precision and recall. The misclassification rate was 0.015, indicating that the model was very effective in differentiating between hypothyroid and non-hypothyroid cases. The SVM model showed good performance in predicting hypothyroidism with high accuracy.

	Misclassification Rate	F1 score	Precision	Recall
SVM	1.5%	0.829	0.895	0.829

Bagging/Random Forest Results

By using bagging and random forest analysis: we arrive at the following results:

	Accuracy	Misclassification Rate	F1 score	Precision	Recall
Bagging	0.9996	0.3%	0.997	1	0.993
RF tree	0.997	0.29%	0.969	0.973	0.966

By comparing the two methods: compared with random forest (0.9971), the accuracy of the Bagging model (0.9997) is slightly higher. Bagging has a higher sensitivity (0.99324) than Random Forest (0.96622), indicating that it is better at identifying true positives. Both models have high specificity, 1.00000 for Bagging and 0.99864 for Random Forest. Bagging has perfect accuracy (1.00000), while Random Forest has slightly lower accuracy (0.97279). Bagging has a slightly higher F1 score (0.99661) compared to Random Forest (0.96949), indicating better overall performance in terms of precision and recall. Bagging has a lower misclassification rate (0.0003237) compared to Random Forest (0.0029136). While both models perform well,

Bagging appears to have a slight advantage over Random Forest in terms of overall performance metrics.

Important Variable and Variable Relationships

Based on the output of the subset selection process, we can summarize the important variables and model sizes for each method (see appendix).

What we ended up with was that TSH, TT4, and T4U were consistently found to be important across all three methods, suggesting that they are key predictors of hypothyroidism.

Based on our correlation plot, we find Age and TSH has positive correlation, suggesting that TSH levels tend to be higher with increasing age, Higher TSH levels will tend to correspond with lower TT4 levels, which is a common clinical finding in hypothyroidism when TSH is elevated when thyroid hormones (like TT4) are low. TSH and FTI have negative correlation, as higher levels of TSH are often found in conjunction with lower levels of free thyroid hormones. Also from our Chi-Squared test results between those Categorical Variables, we find patients querying about hypothyroidism, patients querying about hyperthyroidism are related to being on thyroxine ($p < 0.05$).

Disccsion

With all the analytical methods we performed, we were able to compare models based on multiple criteria to determine the best model for predicting hypothyroidism. When choosing the best model, we typically consider: Accuracy, Accuracy, Recall (sensitivity), F1 Score, Misclassification Rate, and Model Complexity.

PCA shows some of the variables that have the greatest influence on the principal components, which helps with dimensionality reduction but does not directly predict the outcome. LDA provides high performance on all metrics, indicating that the dataset satisfies the assumptions of LDA well. QDA performed very poorly, giving a high misclassification rate. SVM provides high precision and reasonable recall, but does not perform as well as LDA. Both Bagging and Random Forest show very high accuracy and F1 scores, with Bagging slightly better than Random Forest. Considering all the results, Bagging is the best model as it gives extremely high accuracy (almost perfect), lowest misclassification rate, highest F1 score, perfect precision and extremely high recall. This means it is almost perfect at identifying true positive cases and will

not label negative cases as positive. Although bagging models are slightly more complex than Random Forest trees or LDA, their superior performance metrics justify the increased complexity**.**

Conclusion

In conclusion, hypothyroidism is a disease with profound health consequences characterized by insufficient production of thyroid hormone. Our principal component analysis (PCA) reveals key variables affecting the dataset, aiding in dimensionality reduction. Linear discriminant analysis (LDA) gave better performance metrics. In sharp contrast, quadratic discriminant analysis (QDA) performs poorly, and its incompatibility with the underlying data assumptions becomes apparent through high misclassification rates. Support vector machine (SVM) models also gave better performance, especially in minimizing false positives.Also, Bagging outperformed all models, with near-perfect accuracy, extremely low misclassification rates, and unparalleled F1 scores.

In summary, our work highlights the superiority of the Bagging model in predicting hypothyroidism, and through subset selection, we identify TSH, TT4, and T4U as key predictors of hypothyroidism, reaffirming their clinical significance. Furthermore, our correlation analysis revealed an interaction between age and TSH levels, as well as their inverse association with TT4 and FTI, that resonates with clinical patterns observed in hypothyroidism.

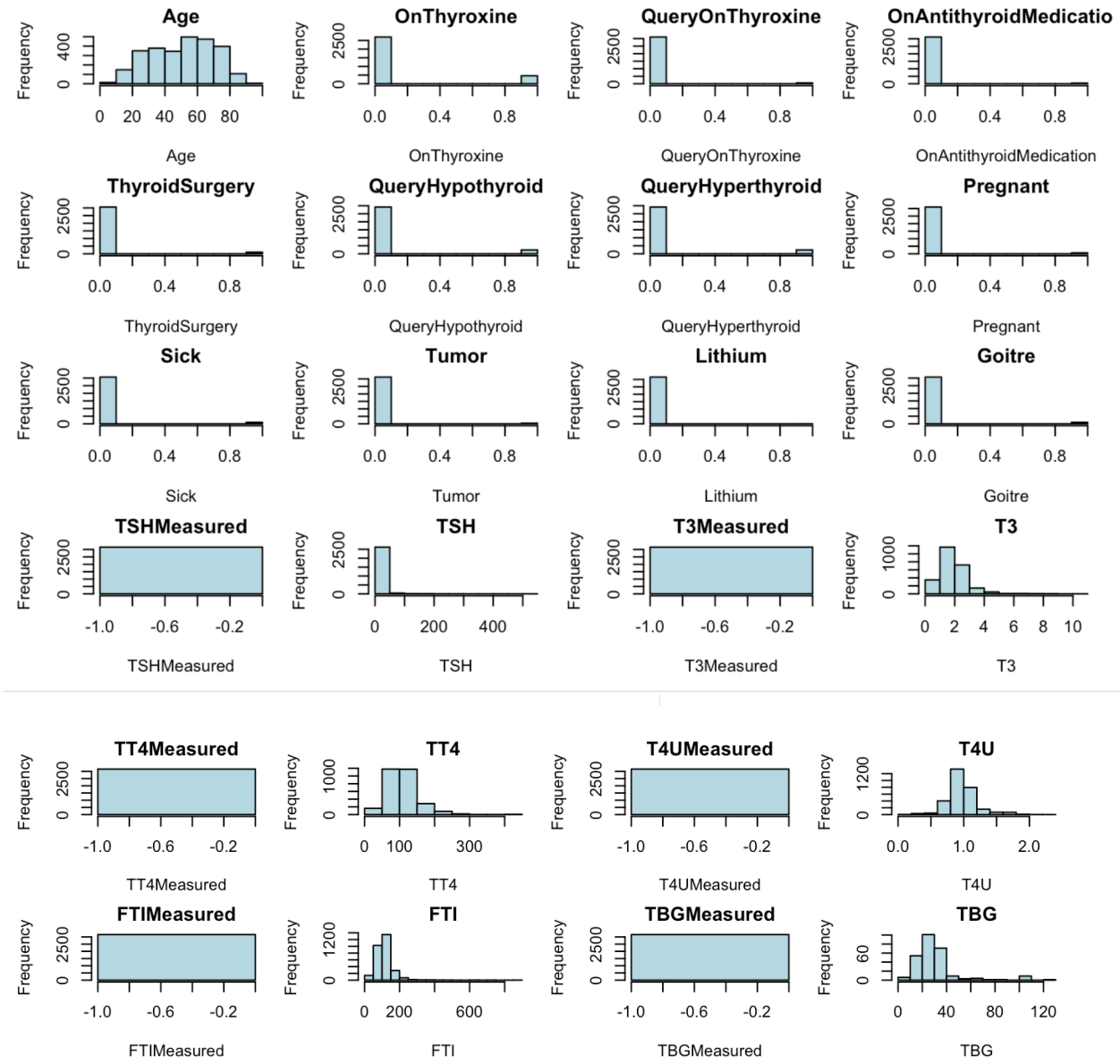
Refrence

https://www.niddk.nih.gov/health-information/endocrine-diseases/hypothyroidism#symptoms

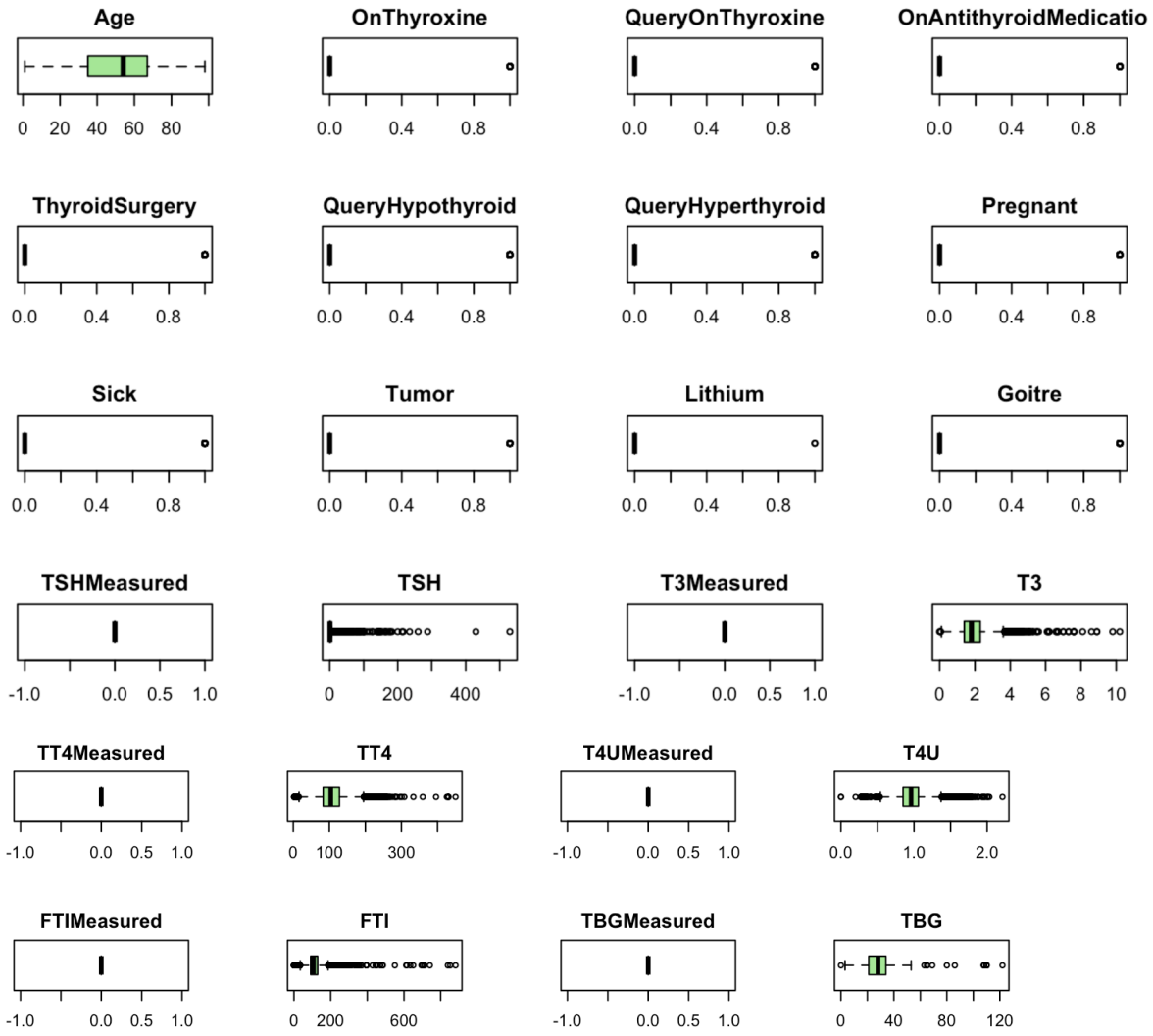
Dataset link(Use hypothyroid.data)

http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease

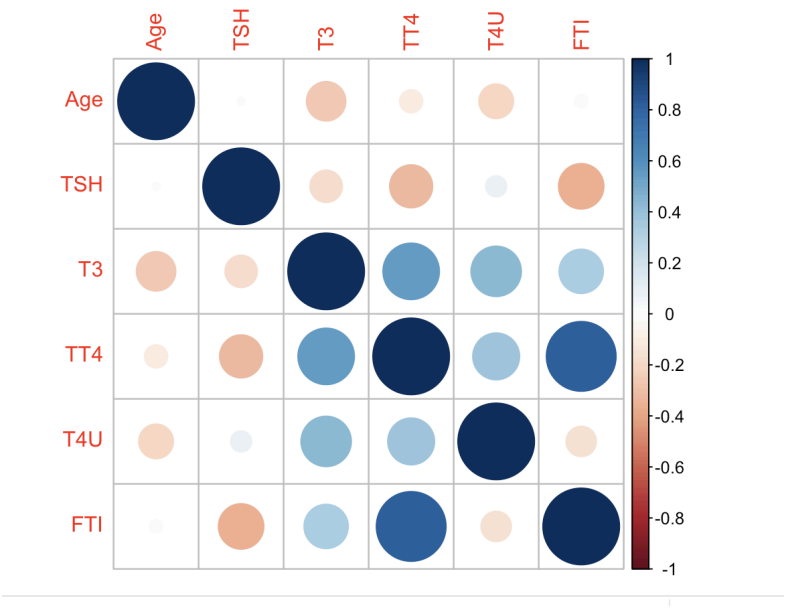
APPENDIX



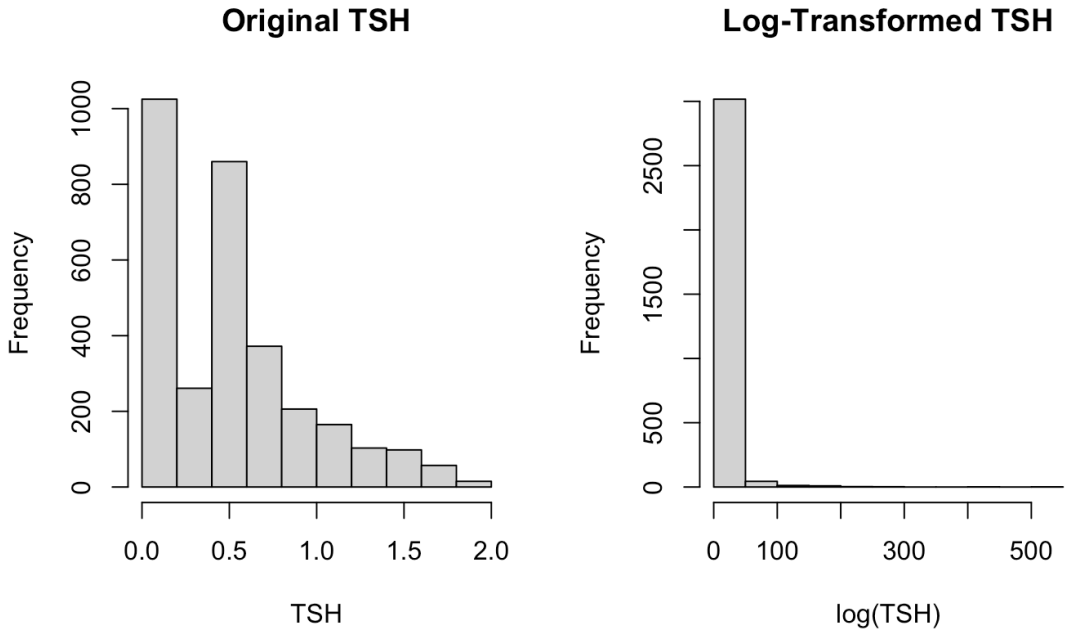
Histograms for All Continuous Variables

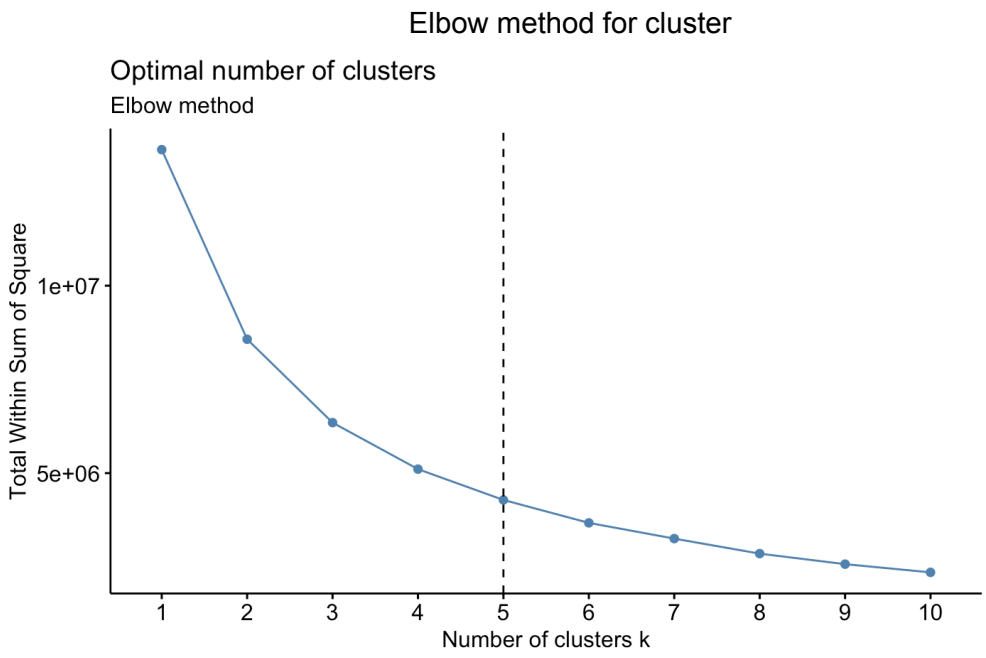


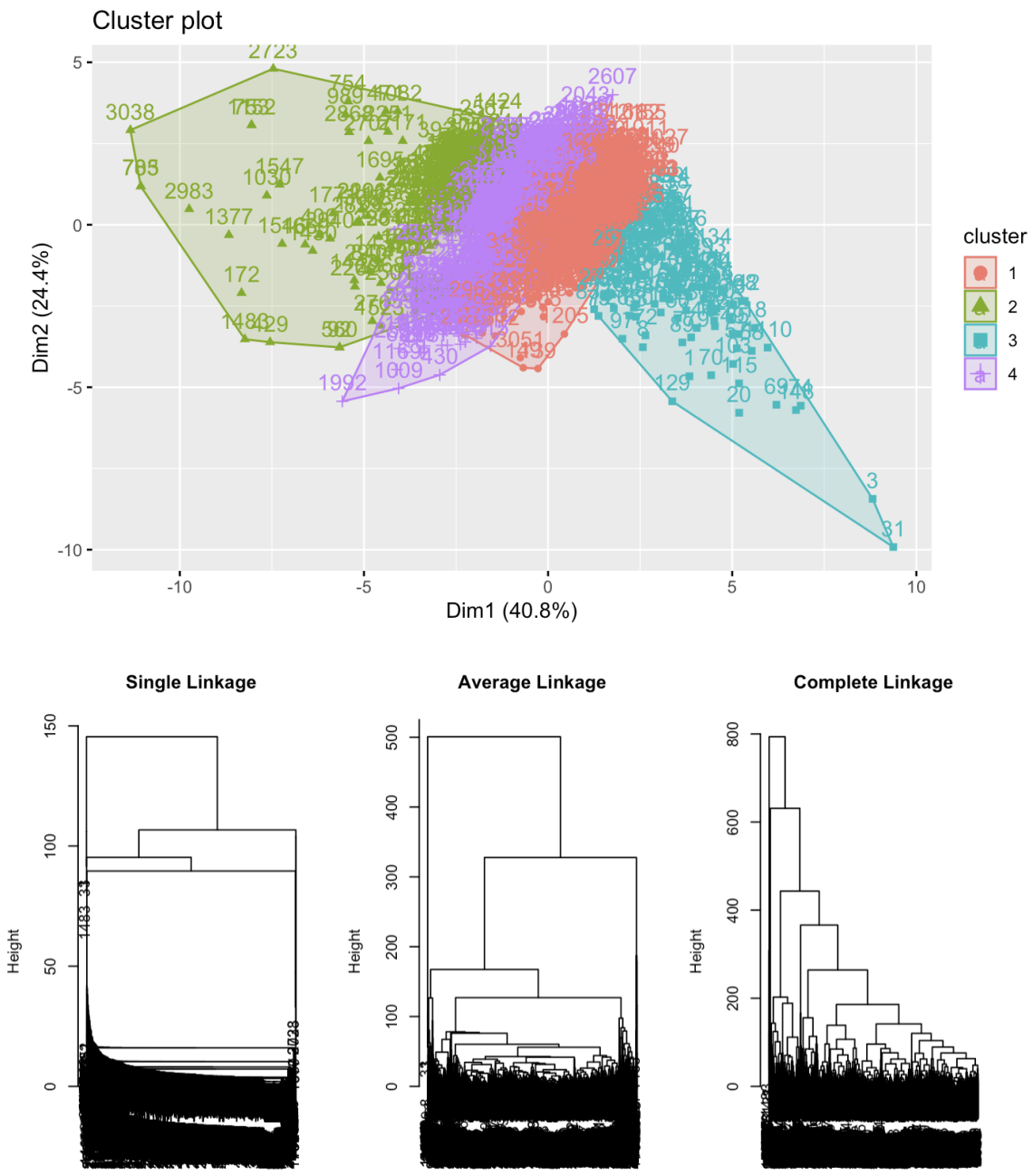
Boxplots for Identifying Outliers



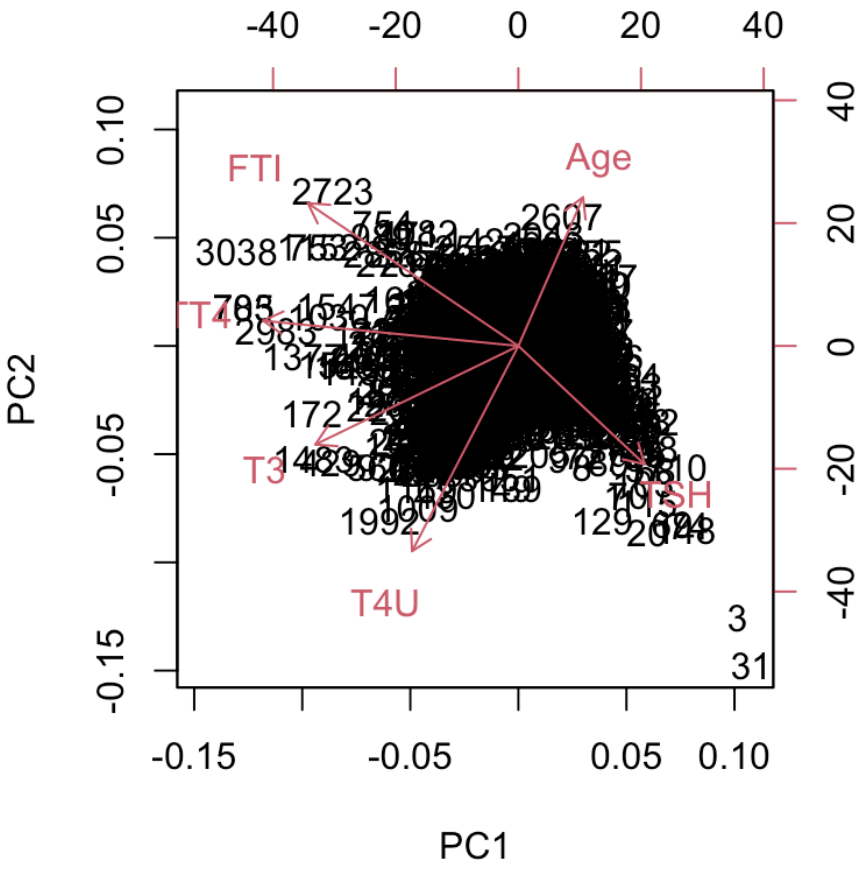
Correlation between continuous variable




Elbow method for cluster
Optimal number of clusters
Elbow method
Total Within Sum of Square
1e+07
5e+06
1
2
3
4
5
6
7
8
9
10
Number of clusters k



Linkage plot and cluster plot

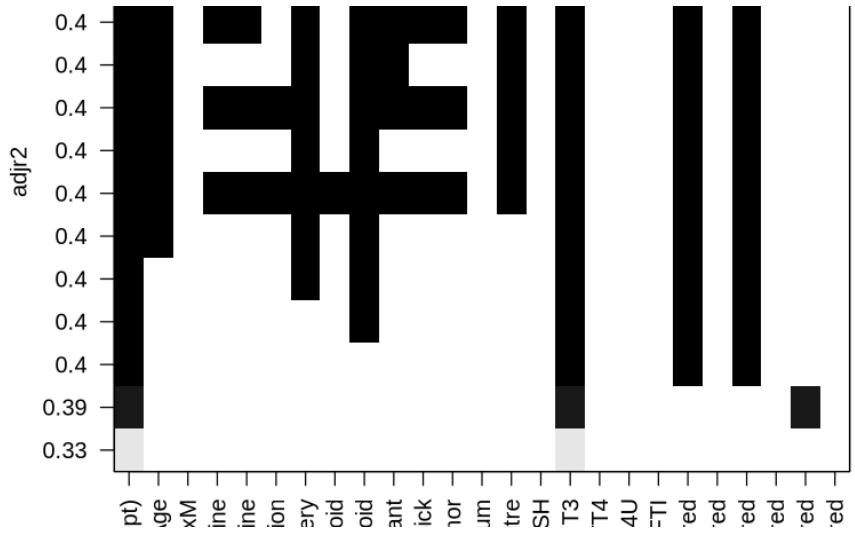


Biplot for PCA

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5    PC6
Standard deviation     1.606 1.2203 0.9330 0.77181 0.66598 0.1450
Proportion of Variance 0.430 0.2482 0.1451 0.09928 0.07392 0.0035
Cumulative Proportion  0.430 0.6782 0.8233 0.92257 0.99650 1.0000
            PC1        PC2         PC3        PC4         PC5          PC6
Age  0.1357504  0.4571563 -0.83305511  0.1225175 -0.25209916 -0.005883162
TSH  0.3890904 -0.3124122 -0.25336590 -0.8249393  0.07926814  0.002166452
T3  -0.4410761 -0.3622975 -0.04587381 -0.1343995 -0.80857870  0.015007093
TT4 -0.5842481  0.0394477 -0.19873585 -0.1992888  0.33274421 -0.683496384
T4U -0.1898673 -0.6387218 -0.44706903  0.3258279  0.36726365  0.339216374
FTI -0.5081701  0.3906732  0.01881046 -0.3748621  0.17539030  0.646142729

Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6
Standard deviation     1.5650 1.2089 0.9380 0.8516 0.67728 0.15964
Proportion of Variance 0.4082 0.2436 0.1467 0.1209 0.07645 0.00425
Cumulative Proportion  0.4082 0.6518 0.7984 0.9193 0.99575 1.00000
NULL
            PC1         PC2        PC3         PC4         PC5          PC6
Age  0.1515604  0.45040670 0.79091031  0.32819740 -0.20216536 -0.006305436
TSH  0.2960203 -0.35874896 0.44524053 -0.76152448 -0.07106811 -0.021490052
T3  -0.4758914 -0.29866372 0.02952171  0.05010832 -0.82518117 -0.004588170
TT4 -0.5980279  0.07650838 0.21160942 -0.15531011  0.31913239 -0.682468968
T4U -0.2490303 -0.62164618 0.35864365  0.35797733  0.40114993  0.365849202
FTI -0.4932296  0.43236034 0.04411805 -0.39697241  0.10192044  0.632351040
```

Original PCA data and transform data



Best subset selection model

	Cp <int>	bic <int>	adjr2 <int>
Best Subset selection	8	4	11
Forward Stepwise	9	7	12
Backward Stepwise	8	4	11

$cp_min_model_size 7

$cp_min_variables '(Intercept)' · 'Age' · 'ThyroidSurgery' · 'QueryHyperthyroid' · 'Goitre' · 'TSH' · 'TT4' · 'T4U'

$bic_min_model_size 3

$bic_min_variables '(Intercept)' · 'TSH' · 'TT4' · 'T4U'

$adjr2_max_model_size 12

$adjr2_max_variables '(Intercept)' · 'Age' · 'OnThyroxine' · 'ThyroidSurgery' · 'QueryHyperthyroid' · 'Pregnant' · 'Tumor' · 'Goitre' · 'TSH' · 'TT4' · 'T4U'

$cp_min_model_size 8

$cp_min_variables '(Intercept)' · 'Age' · 'ThyroidSurgery' · 'QueryHyperthyroid' · 'Goitre' · 'TSH' · 'TT4' · 'T4U' · 'FTI'

$bic_min_model_size 2

$bic_min_variables '(Intercept)' · 'TSH' · 'FTI'

$adjr2_max_model_size 12

$adjr2_max_variables '(Intercept)' · 'Age' · 'OnThyroxine' · 'ThyroidSurgery' · 'QueryHyperthyroid' · 'Pregnant' · 'Tumor' · 'Goitre' · 'TSH' · 'TT4' · 'T4U' · 'FTI'

$cp_min_model_size 7

$cp_min_variables '(Intercept)' · 'Age' · 'ThyroidSurgery' · 'QueryHyperthyroid' · 'Goitre' · 'TSH' · 'TT4' · 'T4U'

$bic_min_model_size 3

$bic_min_variables '(Intercept)' · 'TSH' · 'TT4' · 'T4U'

$adjr2_max_model_size 12

$adjr2_max_variables '(Intercept)' · 'Age' · 'OnThyroxine' · 'ThyroidSurgery' · 'QueryHyperthyroid' · 'Pregnant' · 'Tumor' · 'Goitre' · 'TSH' · 'TT4' · 'T4U'