MEDI504B_ML_Project

Jacob and Hanwei

2023-02-08

1. Introduction and Background

1.1 Introduction

Our overarching goal for this project is to produce a machine learning (ML) model which can accurately predict polycystic ovary syndrome (PCOS) status (presence or absence of disease). PCOS is an endocrine (hormonal) disorder that affects females of a reproductive age. Given the widespread nature of this condition and the troubling symptoms which accompany it, including infertility, it would be helpful for physicians to be able to predict individuals more likely to experience PCOS thereby enabling them to therapeutically intervene and provide care and support in a timely manner.

2. Objectives

- Developing a model which uses physical and clinical predictors to diagnose PCOS status.
- Try different models, and validate and select the model which most accurately predicts PCOS status. Therefore our overarching research question is: what factor or factors predict PCOS status with the highest accuracy?

3. Methods

3.1 Splitting our dataset into training and validation datasets

- Following our EDA (please refer to EDA section), we start by splitting our dataset into the training set and the validation set, followed by building a simple model aimed at using our data to find factors that predict PCOS status (our outcome variable)
- To ensure reproducibility, throughout this project we always set the seed to "504" whenever set.seed() was used.

3.2 Building logistic regression models

We first attempted to model our data with PCOS status as the outcome variable using three different logistic regression models: (1) Model 1 included all the features in our dataset, (2) Model 2 only includes physiological variables we think are of interest in PCOS, namely the hormone measurements (n=5 different types of hormones), and finally (3) Model 3 includes the n=5 the hormone measurement levels we in our dataset, and in addition to these features, the clinical symptoms that have been known to be associated with PCOS: weight gain, hair growth + skin darkening, hair loss and pimples. Of our three logistic regression models, Model 3 had the lowest AIC score of 566.45. We can also use the likelihood ratio test to compare our models.

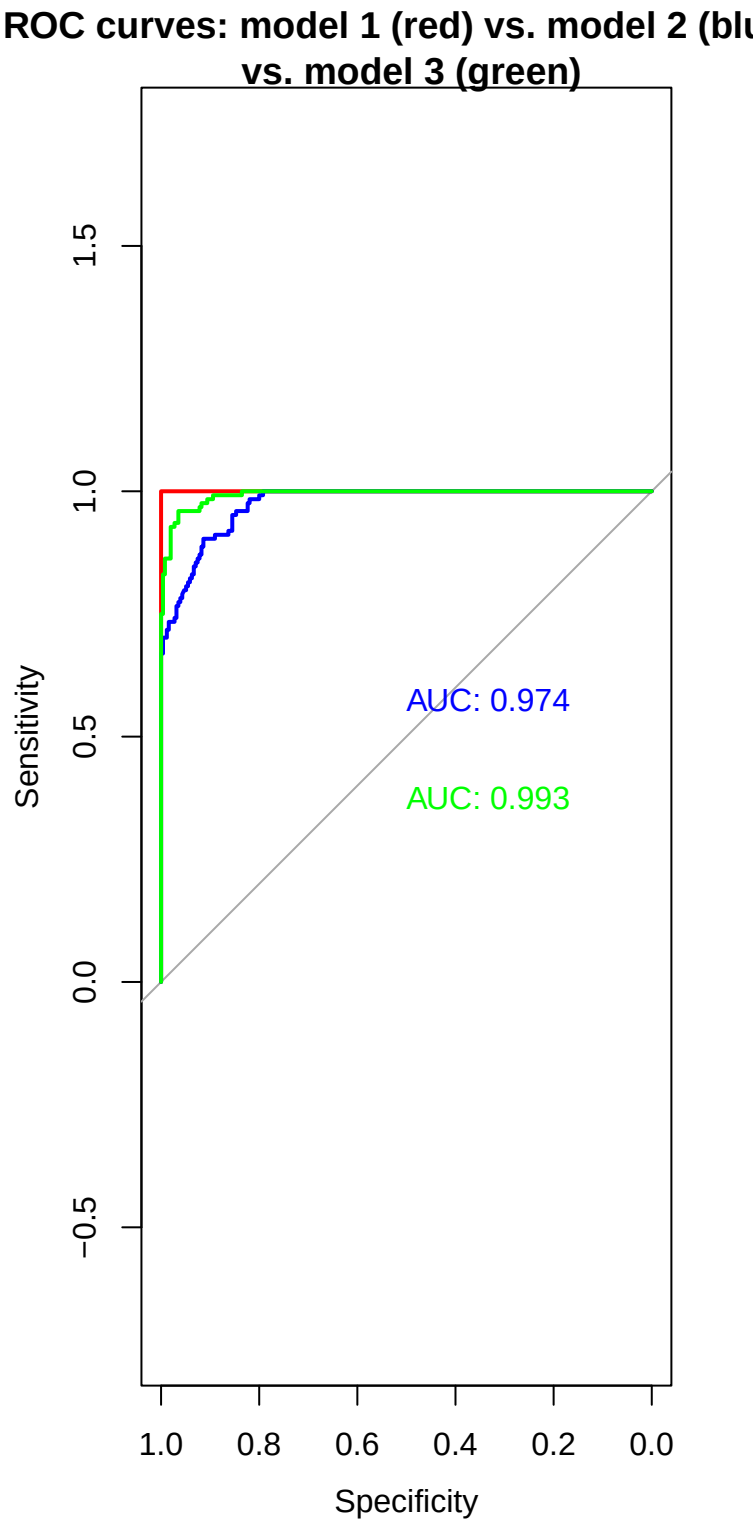
3.3 Model cross-validation (using the caret package)

- We performed model cross-validation on our three logistic regression models and generated a summary of the cross-validation results.

```
##             Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## model1 0.3289474 0.6710526 0.6710526 0.6044211 0.6710526 0.6800000    0
## model2 0.4605263 0.5000000 0.5263158 0.5383626 0.5827193 0.6266667    0
## model3 0.5000000 0.5197368 0.5657895 0.5886784 0.6093860 0.7763158    0
```

3.4 Assessing model classification performance

- Assessing model performance using ROC curves


ROC curves: model 1 (red) vs. model 2 (blu
vs. model 3 (green)
Sensitivity
1.5
1.0
0.5
0.0
−0.5
AUC: 0.974
AUC: 0.993
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

Penalized logistic regression

1.4 Modelling our data using Trees and Forests

Let's select only the predictor columns we decided we wanted to include earlier. From logistic model 3, here is the list of covariates: * pcos_y_n * i_beta_hcg_m_iu_m_l * fsh_m_iu_m_l * lh_m_iu_m_l * tsh_m_iu_l * amh_ng_m_l * weight_gain_y_n * hair_growth_y_n * skin_darkening_y_n * hair_loss_y_n * pimples_y_n

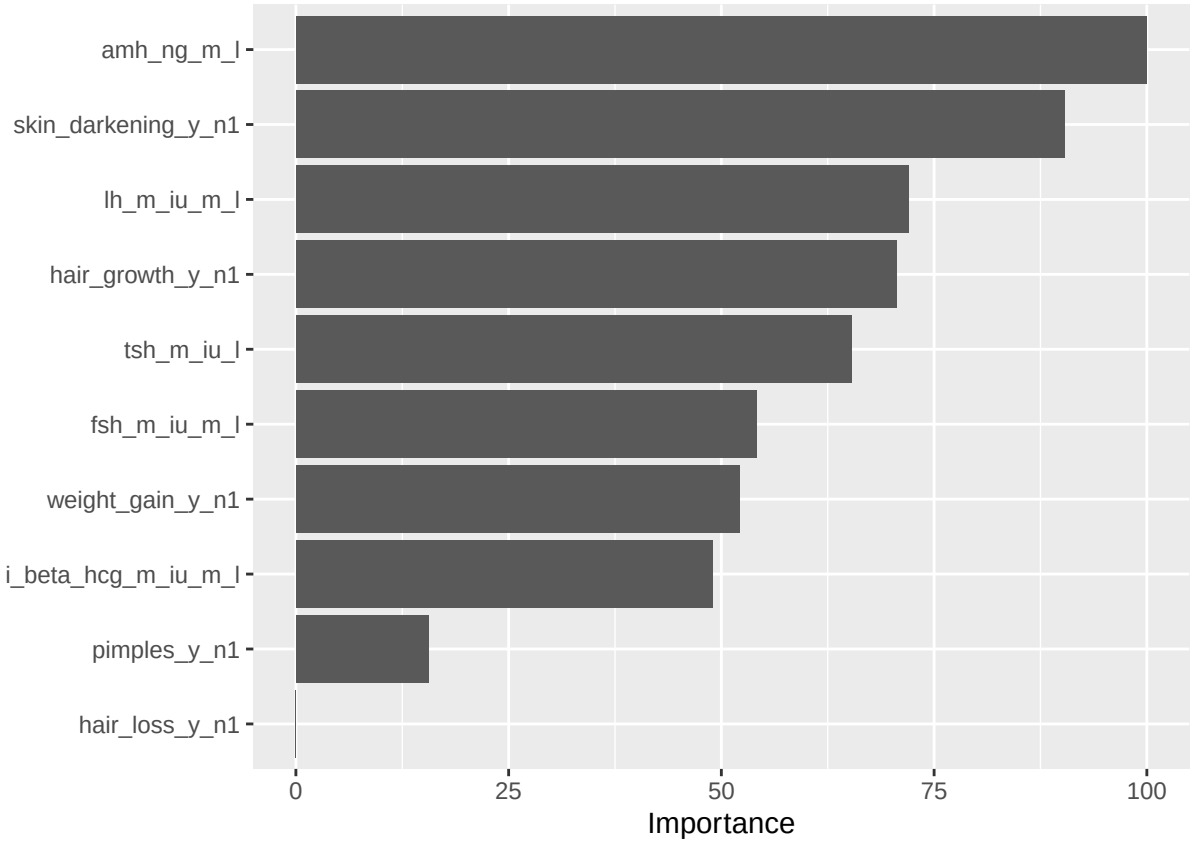
```
train.data.covariates <- train.data %>%
  select(
    pcos_y_n,
    i_beta_hcg_m_iu_m_l,
    fsh_m_iu_m_l ,
    lh_m_iu_m_l ,
    tsh_m_iu_l ,
    amh_ng_m_l ,
    weight_gain_y_n ,
    hair_growth_y_n,
    skin_darkening_y_n,
    hair_loss_y_n ,
    pimples_y_n
  ) %>% mutate(
  amh_ng_m_l = as.numeric(amh_ng_m_l),
  pcos_labelled = as.factor(ifelse(pcos_y_n==1, "yes", "no")))
```

```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

Random Forest – note that I've had to remove one missing ob from amh_ng_m_l for now.

RF Variable Importance

```
vip::vip(caret_rf)
```



XGboost

```
set.seed(123)
xgb_grid_1  <-  expand.grid(
                  nrounds = 50,
                  eta = c(0.03),
                  max_depth = 1,
                  gamma = 0,
                  colsample_bytree = 0.6,
                  min_child_weight = 1,
                  subsample = 0.5
                )

caret_xgb <- caret::train(pcos_labelled ~., data = select(train.data.covariates, -(pcos_y_n)),
                          method = "xgbTree",
                          metric = "ROC",
                          tuneGrid=xgb_grid_1,
                          na.action = na.pass,
                          trControl = trainControl(method = "cv", number = 5,classProbs = T, summaryFunc
caret_xgb
```

```
## eXtreme Gradient Boosting
##
```

```
## 379 samples
##  10 predictor
##   2 classes: 'no', 'yes'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 303, 304, 303, 303, 303
## Resampling results:
##
##   ROC        Sens       Spec
##   0.8800556  0.9372549  0.525
##
## Tuning parameter 'nrounds' was held constant at a value of 50
## Tuning
##  held constant at a value of 1
## Tuning parameter 'subsample' was held
##  constant at a value of 0.5
```

Comparison code for later

4. Results

4a. Exploratory Data Analysis

This section contains the steps and output for the EDA performed on the PCOS dataset. The section proceeds sequentially with each step of EDA. Please refer to bullet points, figure titles and figure captions for more details. The first steps in our EDA involved getting a high-level overview of our dataset and determining the dimension of our data: n=541 rows and n=45 columns.

- Further checking of the dataset reveals an addtional column (column 45)
- This column does not contain any useful information and is not one of our 44 features, therefore the column was removed
- We find that our dataset has the following variables: Sl. No, Patient File No., PCOS (Y/N), Age (yrs), Weight (Kg), Height(Cm), BMI, Blood Group, Pulse rate(bpm), RR(breaths/min), Hb(g/dl), Cycle(R/I), Cycle length(days), Marriage Status (Yrs), Pregnant(Y/N), No. of aborptions, I beta-HCG(mIU/mL), II beta-HCG(mIU/mL), FSH(mIU/mL), LH(mIU/mL), FSH/LH, Hip(inch), Waist(inch), Waist:Hip Ratio, TSH (mIU/L), AMH(ng/mL), PRL(ng/mL), Vit D3 (ng/mL), PRG(ng/mL), RBS(mg/dl), Weight gain(Y/N), hair growth(Y/N), Skin darkening (Y/N), Hair loss(Y/N), Pimples(Y/N), Fast food (Y/N), Reg.Exercise(Y/N), BP _Systolic (mmHg), BP _Diastolic (mmHg), Follicle No. (L), Follicle No. (R), Avg. F size (L) (mm), Avg. F size (R) (mm), Endometrium (mm)

Data Wrangling

Next, we formatted the data to ensure the variables are the appropriate class type, this will enable us to perform our EDA. Specifically we converted binary variables (1 or 0) into the character class type.

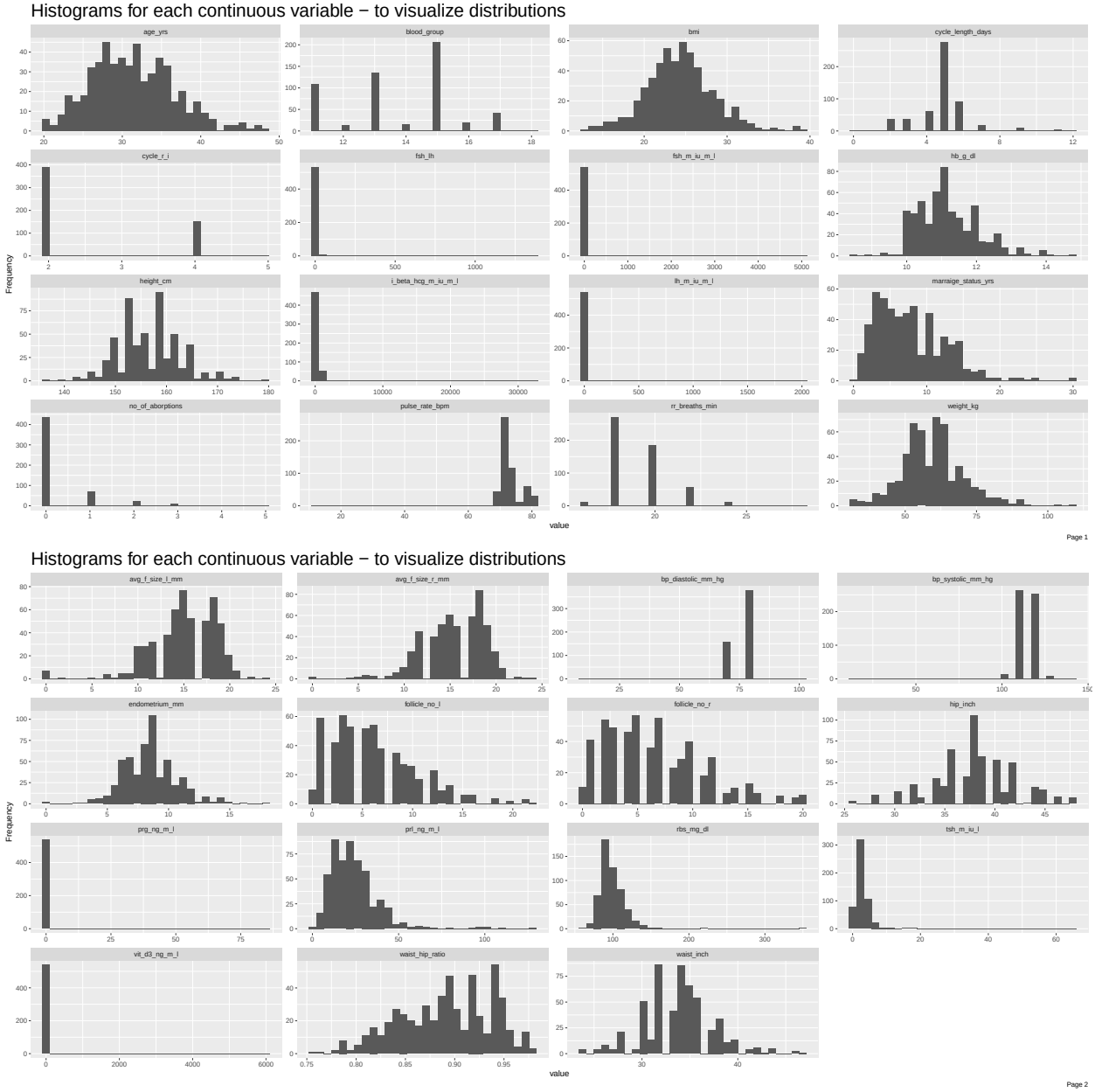
Missing Values

Next in our EDA we sought to identify missing values in our dataset. Here, we see a plot showing our variables and the percentage of missing rows per variable.

The plot shows the missing values in our dataset. Our EDA identified some missing values for two variables: fast food and marriage status. The analysis indicates that only 0.18% of the rows for these variables are missing. Therefore, as this is below the generally used threshold of 5 %, we will simply ignore these missing values for our subsequent analyses.

Univariate distributions for continuous variables

To get a sense of the variation in our dataset, we plotted histograms for each continuous variable using the plot_histogram() from the DataExplorer Package.



- We see that the age distribution in our dataset reveals most individuals are in the range of 20 to 40 years. No individuals in the dataset are younger than 20 or older than 50.
- As we would expect, the BMI values follow an approximately normal distribution.

- The most common blood type we observe is O+, which is consistent with the fact that O+ is the most frequent bloodtype globally.
- The most common cycle length is 5 days.
- The endometrium thickness data suggests there are two most most common thickness values (two clear peaks in the distribution).

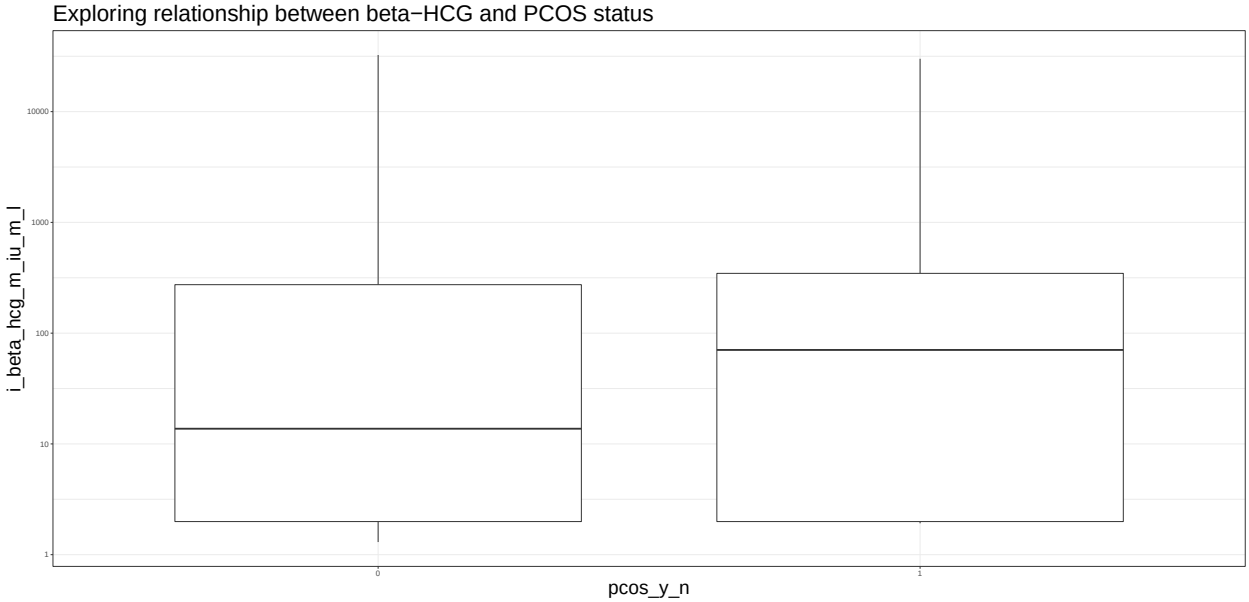
Bivariate associations

- From this correlation plot, we find that several continuous variables do co-vary with one another.
- Specifically, as we would expect, we find a positive correlation between the variables waist and hip (in inches) with weight. We find the same positive correlation for BMI.
- Another obvious correlation we observe is that between age (in years) and marriage (in years)
-

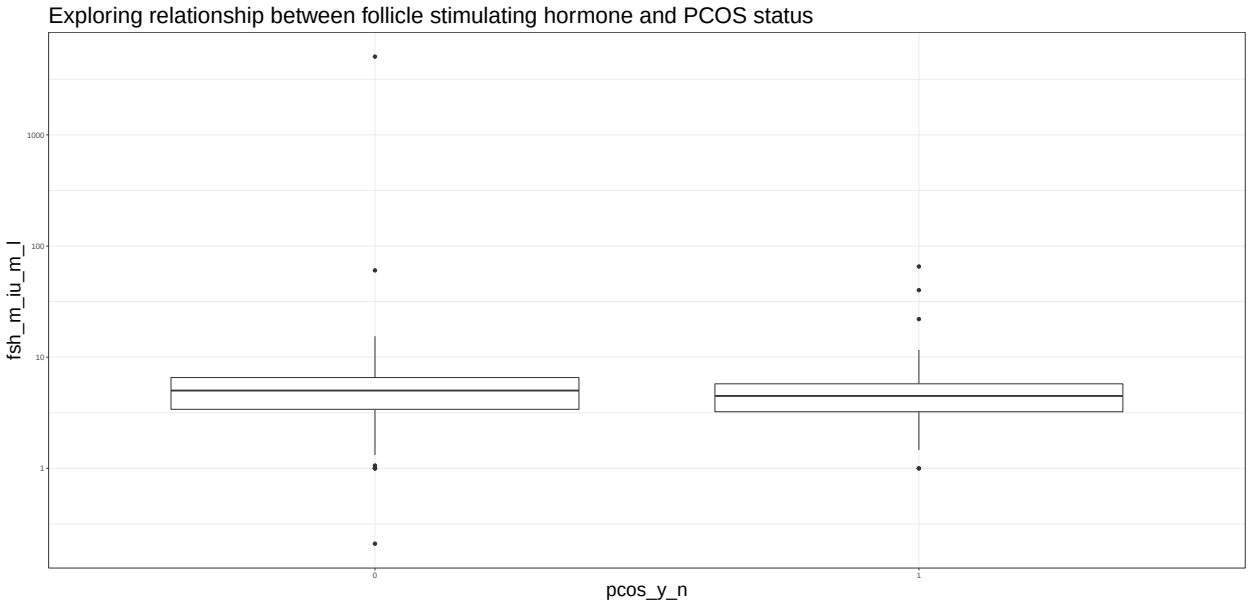
- These bar plots show us the relative frequencies of the categorical features in our dataset.
- Further the bar plots have been coloured to show frequencies according to our outcome variable of PCOS status.
- We can see, for instance, the proportion of individuals with weight gain and a positive PCOS status is greater than the proportion of individuals with no weight gain and a positive PCOS status. We observe similar associations for hair growth and skin darkening. We also observe a similar but much weaker association for the categorical variables of hair loss and pimples. Interestingly, we observe the inverse trend for the categorical variable for fast food consumption.
- The bar plots for regular exercise and pregnancy do not appear to show a signfiicant difference in proportion of positive PCOS cases, but we would need to perform a statistical analyses (such as an unpaired t-test) to know for sure.
- Following the global EDA of our data, to investigate relationships between our continuous variables and PCOS status (Y/N), we plotted boxplots to visualize any associations between PCOS status and a continuous variable. Please refer to the appendix for these boxplots.

- For this section of the EDA, we looked more closely at the relationships between selected variables and our outcome variable of interest (for this project: yes or no for PCOS), based on how our boxplots looked initially (please refer to Appendix for all boxplots) and on some knowledge gained via our literature survey, we will highlight the relatiionship of select variables with our outcome variable (PCOS status).
- While the etiology for PCOS is not known, my literature survey suggests PCOS is associated with abnormal hormone levels. Thus, to look into this further as part of the EDA, we compared measurements of hormone levels and PCOS status to get a sense of the relationships between these variables.

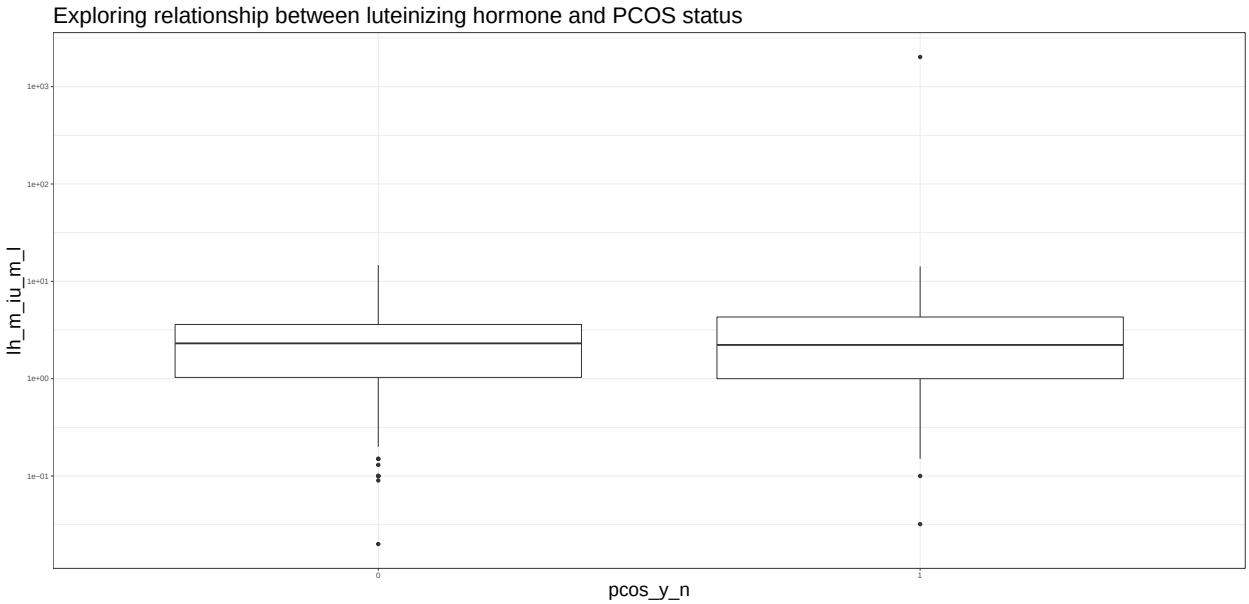
Boxplots looking at relationship of hormones and PCOS status



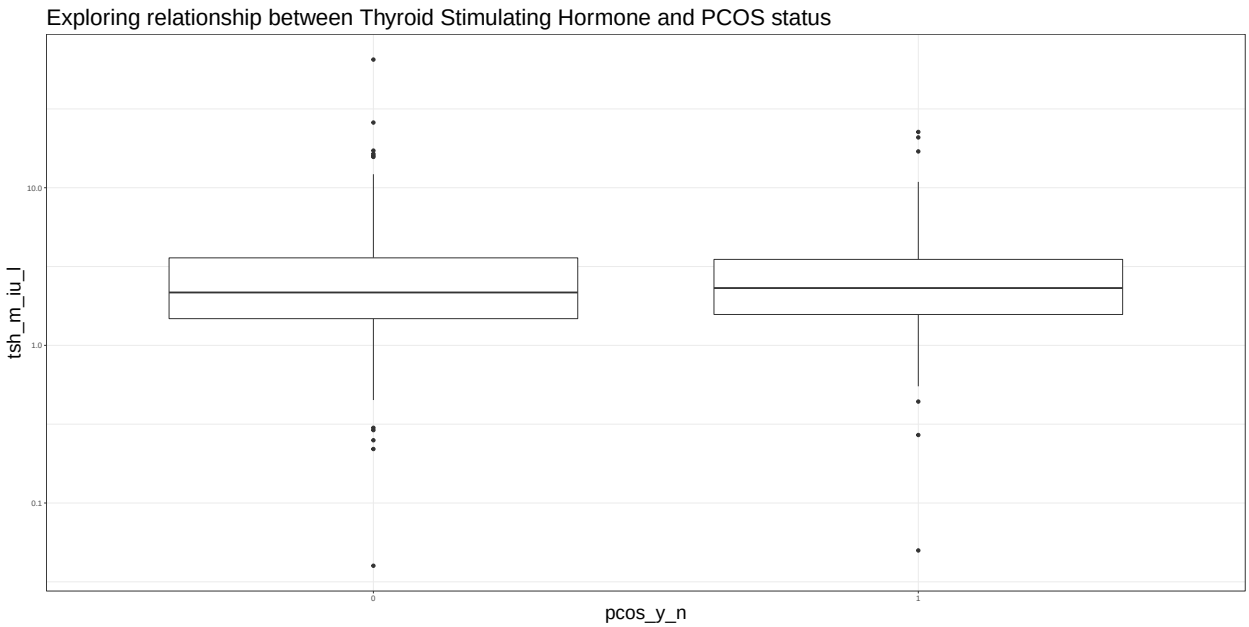
EDA exploring the associationg between beta−HCG hormone and our outocme variable PCOS status



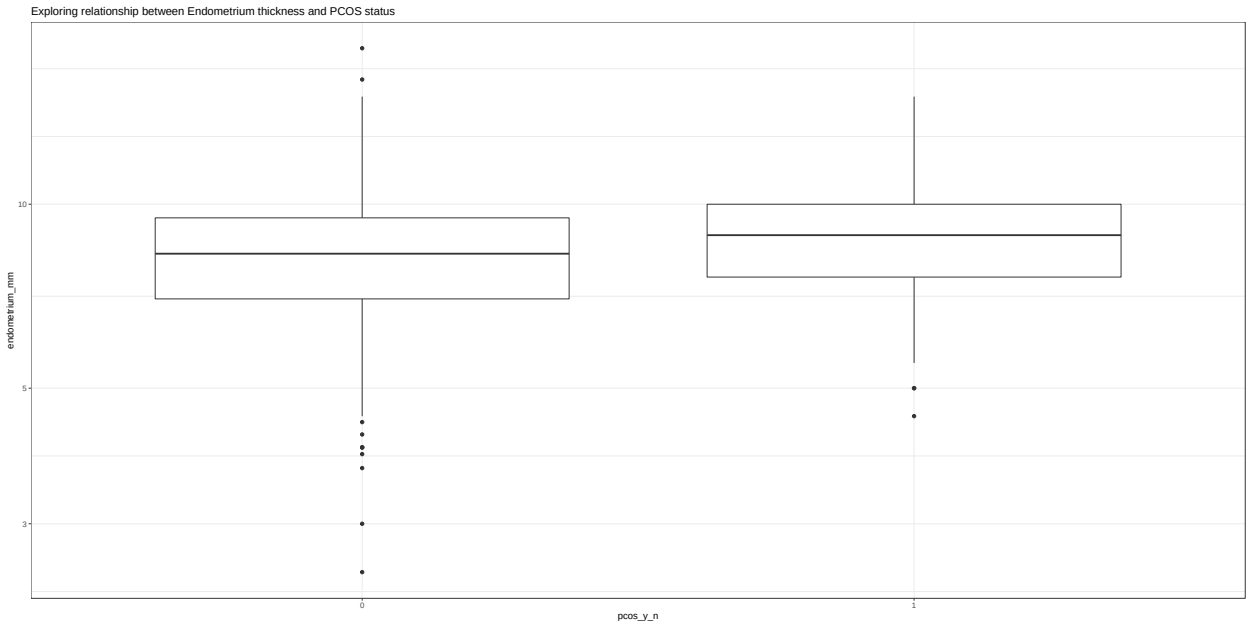
EDA exploring the association between follicle stimulating hormone and our outcome variable PCOS status



EDA exploring the association between Lutenizing hormone and PCOS status.



- From the literature, we know that PCOS diagnosis often involves an ultrasound and endometrium thickness is measured as part of this process. Therefore, as part of our EDA, we also sought to look at the relationship between endometrium thickness and PCOS status.



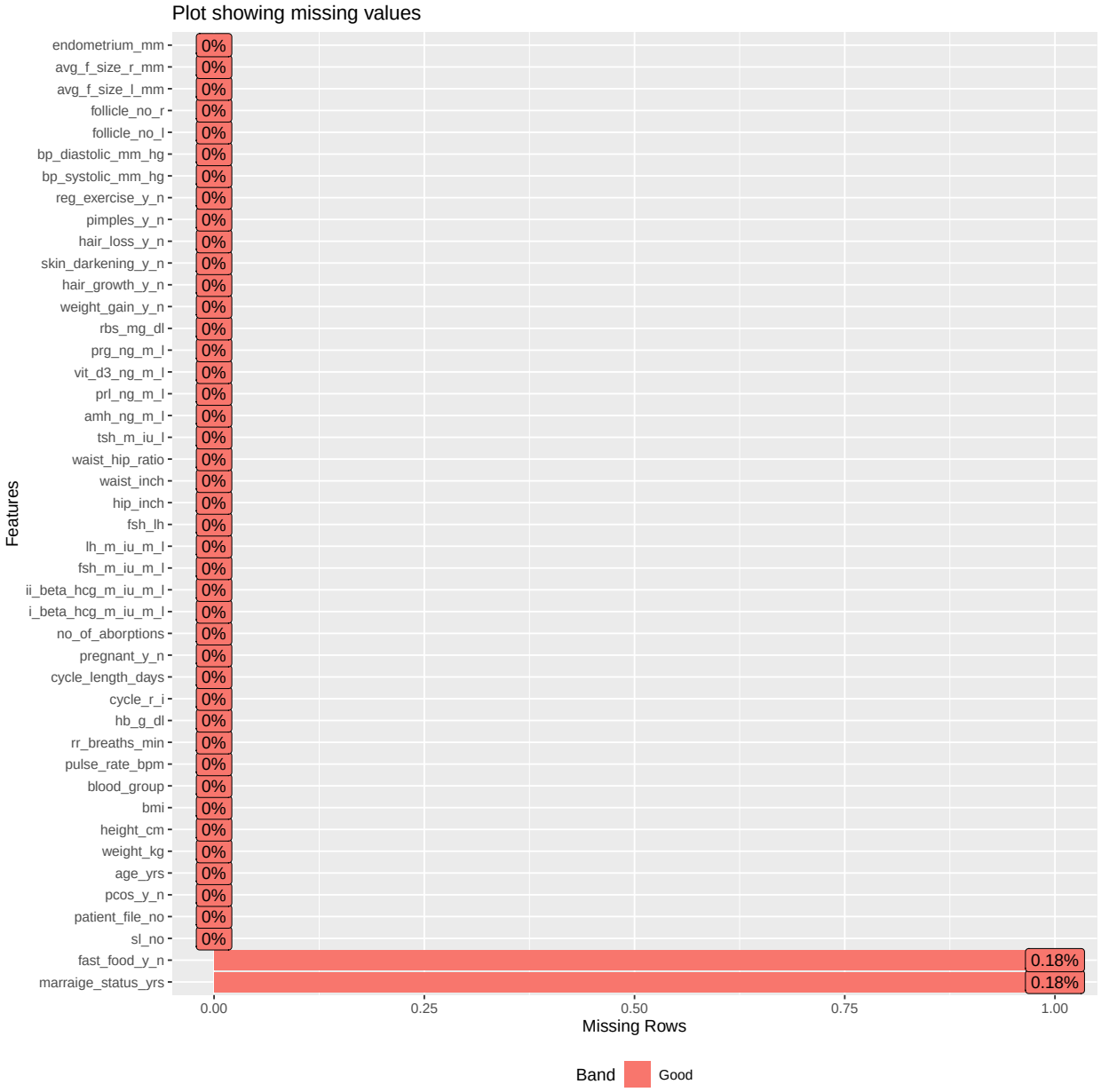
- From this series of boxplots, we did not observe (visually) any significant difference in the association of PCOS status and hormone levels. However, given what we know about the association of PCOS and hormone imbalances from the literature, I think it would still be useful to include these variables when we build our model.
- From the literature[2], there is also evidence that miscarriages are associated with PCOS. Therefore, as part of our EDA, we sought to look at the relationship between PCOS status and number of abortions.

References

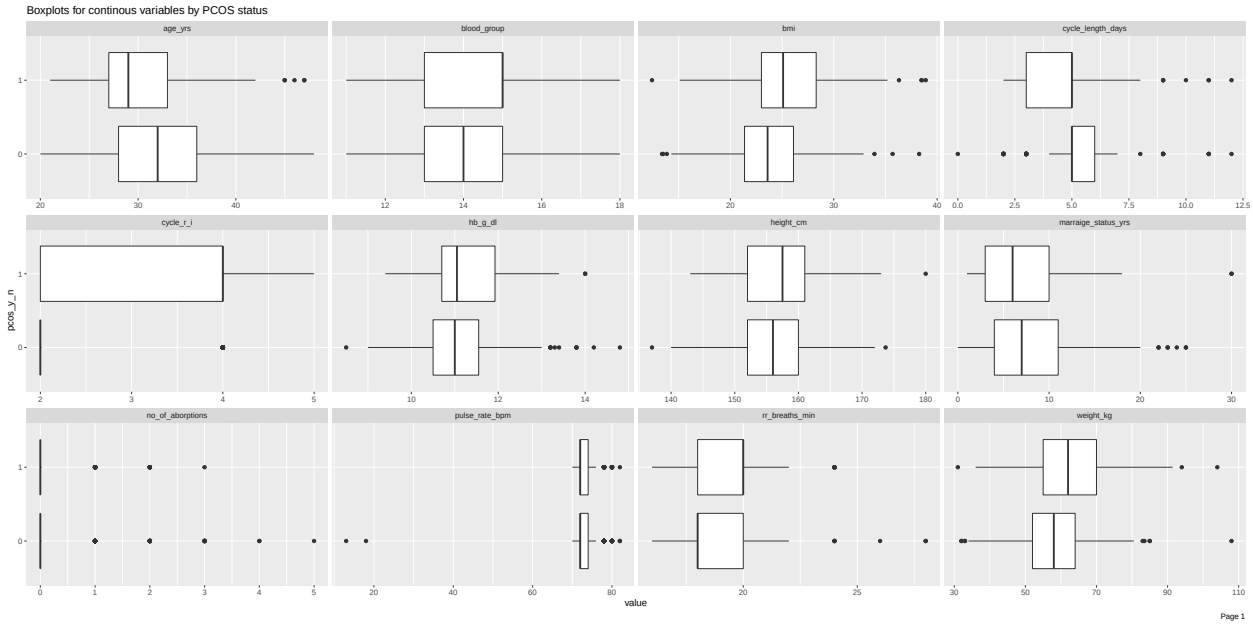
1. R for Data Science by Hadley Wickham (https://r4ds.had.co.nz)
2. Ajmal N, Khan SZ, Shaikh R. Polycystic ovary syndrome (PCOS) and genetic predisposition: A review article. Eur J Obstet Gynecol Reprod Biol X. 2019 Jun 8;3:100060. doi: 10.1016/j.eurox.2019.100060. PMID: 31403134; PMCID: PMC6687436.

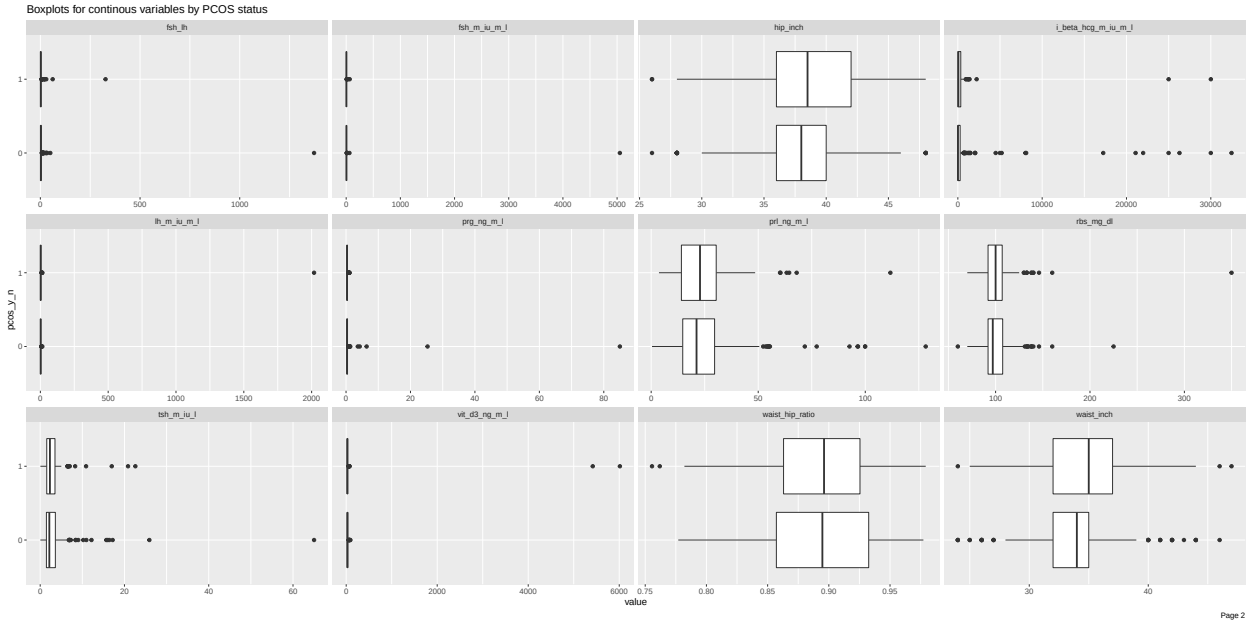
Appendix

- Appendix Plot 1, as part of EDA, showing missing values in our dataset

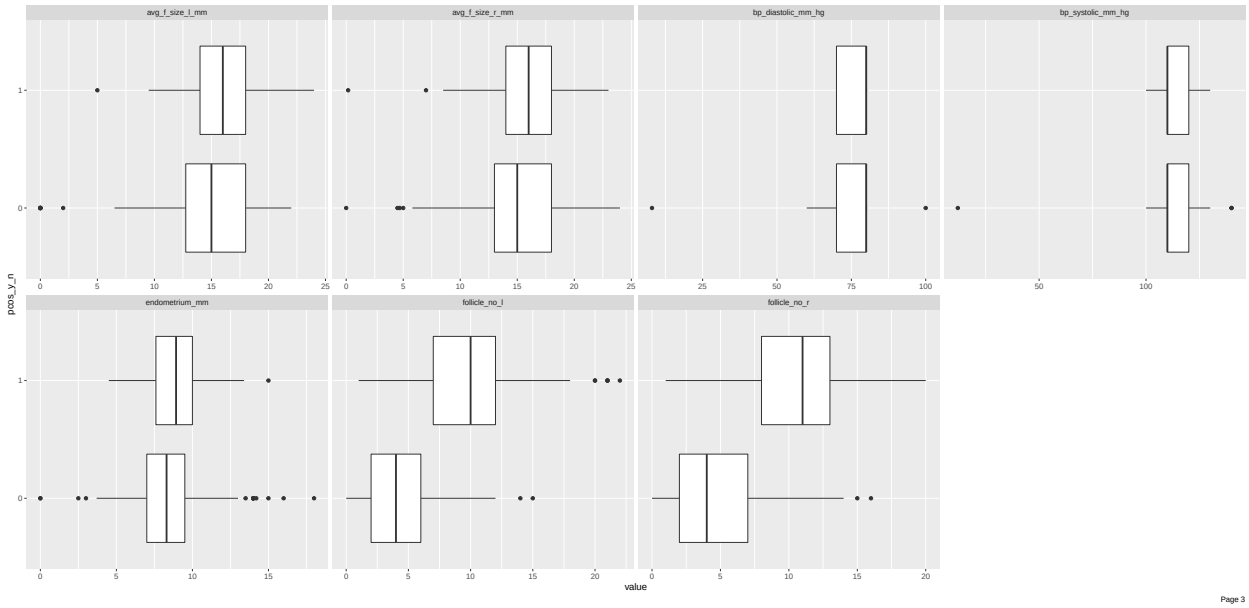

Plot showing missing values
endometrium_mm
avg_f_size_r_mm
avg_f_size_l_mm
follicle_no_r
follicle_no_l
bp_diastolic_mm_hg
bp_systolic_mm_hg
reg_exercise_y_n
pimples_y_n
hair_loss_y_n
skin_darkening_y_n
hair_growth_y_n
weight_gain_y_n
rbs_mg_dl
prg_ng_m_l
vit_d3_ng_m_l
prl_ng_m_l
amh_ng_m_l
tsh_m_iu_l
waist_hip_ratio
waist_inch
hip_inch
fsh_lh
lh_m_iu_m_l
fsh_m_iu_m_l
ii_beta_hcg_m_iu_m_l
i_beta_hcg_m_iu_m_l
no_of_aborptions
pregnant_y_n
cycle_length_days
cycle_r_i
hb_g_dl
rr_breaths_min
pulse_rate_bpm
blood_group
bmi
height_cm
weight_kg
age_yrs
pcos_y_n
patient_file_no
sl_no
fast_food_y_n
marraige_status_yrs
0%
0.18%
Features
0.00
0.25
0.50
0.75
1.00
Missing Rows
Band
Good

- Appendix plot 2, as part of EDA, examining correlations between continuous features
- Appendix Plot 3 showing boxplots generated to investigate potential associations between continuous variables and PCOS status.




Boxplots for continous variables by PCOS status
fsh_lh
fsh_m_iu_m_l
hip_inch
i_beta_hcg_m_iu_m_l
lh_m_iu_m_l
prg_ng_m_l
prl_ng_m_l
rbs_mg_dl
tsh_m_iu_l
vit_d3_ng_m_l
waist_hip_ratio
waist_inch
pcos_y_n
value

Page 2


Boxplots for continous variables by PCOS status
avg_f_size_l_mm
avg_f_size_r_mm
bp_diastolic_mm_hg
bp_systolic_mm_hg
endometrium_mm
follicle_no_l
follicle_no_r
pcos_y_n
value

Page 3