MEDI 504B - Exploratory Data Analysis

Erick Navarro & Timo Tolppa

2023-01-11

Contents

1 Introduction

The goal of this report is to conduct an exploratory data analysis (EDA) of a publicly available dataset of polycystic ovary syndrom (PCOS), a hormonal disorder common among women of reproductive age.

This is the first step of the course project of developing a model to diagnose PCOS. On this deliverable, I will explore the dataset, clean it, and understand its variables.

2 Data Preparation & Pre-processing

2.1 Load libraries and data files

The packages that need to be installed for this exploratory data analysis include 'janitor', 'tidyverse', 'DataExplorer', 'skimr', 'here', 'knitr' and 'readxl.' The names of the columns are cleaned using the janitor package.

```
# Load required libraries
library(tidyverse)
library(here)
library(readxl)
library(janitor)
library(DataExplorer)
library(knitr)
library(skimr)
library(cowplot)

# Load the data file to a data frame
data = read_excel(here("PCOS_data_without_infertility.xlsx"), sheet = "Full_new") %>%
  clean_names()
```

2.2 Data preprocessing

An overview of the data is shown using 'glimpse' and 'skim'.

```
glimpse(data)
```

```
## Rows: 541
## Columns: 45
## $ sl_no                 <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15~
## $ patient_file_no       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15~
## $ pcos_y_n              <dbl> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0~
## $ age_yrs               <dbl> 28, 36, 33, 37, 25, 36, 34, 33, 32, 36, 20, 26, 2~
## $ weight_kg             <dbl> 44.6, 65.0, 68.8, 65.0, 52.0, 74.1, 64.0, 58.5, 4~
## $ height_cm             <dbl> 152.0, 161.5, 165.0, 148.0, 161.0, 165.0, 156.0, ~
## $ bmi                   <dbl> 19.30000, 24.92116, 25.27089, 29.67495, 20.06095,~
## $ blood_group           <dbl> 15, 15, 11, 13, 11, 15, 11, 13, 11, 15, 15, 13, 1~
## $ pulse_rate_bpm        <dbl> 78, 74, 72, 72, 72, 78, 72, 72, 72, 80, 80, 72, 7~
## $ rr_breaths_min        <dbl> 22, 20, 18, 20, 18, 28, 18, 20, 18, 20, 20, 20, 1~
## $ hb_g_dl               <dbl> 10.48, 11.70, 11.80, 12.00, 10.00, 11.20, 10.90, ~
## $ cycle_r_i             <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 4, 2, 2, 4, 2, 2, 2, 2~
## $ cycle_length_days     <dbl> 5, 5, 5, 5, 5, 5, 5, 5, 5, 2, 5, 5, 2, 5, 5, 5, 5~
## $ marraige_status_yrs   <dbl> 7, 11, 10, 4, 1, 8, 2, 13, 8, 4, 4, 3, 7, 15, 9, ~
## $ pregnant_y_n          <dbl> 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1~
## $ no_of_aborptions      <dbl> 0, 0, 0, 0, 0, 0, 0, 2, 1, 0, 2, 1, 0, 0, 0, 0, 0~
## $ i_beta_hcg_m_iu_m_l   <dbl> 1.99, 60.80, 494.08, 1.99, 801.45, 237.97, 1.99, ~
## $ ii_beta_hcg_m_iu_m_l  <chr> "1.99", "1.99", "494.08", "1.99", "801.45", "1.99~
## $ fsh_m_iu_m_l          <dbl> 7.95, 6.73, 5.54, 8.06, 3.98, 3.24, 2.85, 4.86, 3~
## $ lh_m_iu_m_l           <dbl> 3.68, 1.09, 0.88, 2.36, 0.90, 1.07, 0.31, 3.07, 3~
## $ fsh_lh                <dbl> 2.160326, 6.174312, 6.295455, 3.415254, 4.422222,~
## $ hip_inch              <dbl> 36, 38, 40, 42, 37, 44, 39, 44, 39, 40, 39, 39, 4~
## $ waist_inch            <dbl> 30, 32, 36, 36, 30, 38, 33, 38, 35, 38, 35, 33, 4~
## $ waist_hip_ratio       <dbl> 0.8333333, 0.8421053, 0.9000000, 0.8571429, 0.810~
## $ tsh_m_iu_l            <dbl> 0.68, 3.16, 2.54, 16.41, 3.57, 1.60, 1.51, 12.18,~
## $ amh_ng_m_l            <chr> "2.07", "1.53", "6.63", "1.22", "2.26", "6.74", "~
## $ prl_ng_m_l            <dbl> 45.16, 20.09, 10.52, 36.90, 30.09, 16.18, 26.41, ~
## $ vit_d3_ng_m_l         <dbl> 17.10, 61.30, 49.70, 33.40, 43.80, 52.40, 42.70, ~
## $ prg_ng_m_l            <dbl> 0.57, 0.97, 0.36, 0.36, 0.38, 0.30, 0.46, 0.26, 0~
## $ rbs_mg_dl             <dbl> 92, 92, 84, 76, 84, 76, 93, 91, 116, 125, 108, 10~
## $ weight_gain_y_n       <dbl> 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0~
## $ hair_growth_y_n       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0~
## $ skin_darkening_y_n    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0~
## $ hair_loss_y_n         <dbl> 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0~
## $ pimples_y_n           <dbl> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0~
## $ fast_food_y_n         <dbl> 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0~
## $ reg_exercise_y_n      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0~
## $ bp_systolic_mm_hg     <dbl> 110, 120, 120, 120, 120, 110, 120, 120, 120, 110,~
## $ bp_diastolic_mm_hg    <dbl> 80, 70, 80, 70, 80, 70, 80, 80, 80, 80, 80, 80, 8~
## $ follicle_no_l         <dbl> 3, 3, 13, 2, 3, 9, 6, 7, 5, 1, 7, 4, 15, 3, 4, 1,~
## $ follicle_no_r         <dbl> 3, 5, 15, 2, 4, 6, 6, 6, 7, 1, 15, 2, 8, 3, 1, 3,~
## $ avg_f_size_l_mm       <dbl> 18, 15, 18, 15, 16, 16, 15, 15, 17, 14, 17, 18, 2~
## $ avg_f_size_r_mm       <dbl> 18, 14, 20, 14, 14, 20, 16, 18, 17, 17, 20, 19, 2~
## $ endometrium_mm        <dbl> 8.5, 3.7, 10.0, 7.5, 7.0, 8.0, 6.8, 7.1, 4.2, 2.5~
## $ x45                   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
```

```
skim(data)
```

Table 1: Data summary

Name	data

Table 1: Data summary

Number of rows	541
Number of columns	45
Column type frequency:	
character	3
numeric	42
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ii_beta_hcg_m_iu_m_l	0	1	3	8	0	203	0
amh_ng_m_l	0	1	1	5	0	301	0
x45	539	0	1	3	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
sl_no	0	1	271.00	156.32	1.00	136.00	271.00	406.00	541.00	
patient_file_no	0	1	271.00	156.32	1.00	136.00	271.00	406.00	541.00	
pcos_y_n	0	1	0.33	0.47	0.00	0.00	0.00	1.00	1.00	
age_yrs	0	1	31.43	5.41	20.00	28.00	31.00	35.00	48.00	
weight_kg	0	1	59.64	11.03	31.00	52.00	59.00	65.00	108.00	
height_cm	0	1	156.48	6.03	137.00	152.00	156.00	160.00	180.00	
bmi	0	1	24.31	4.06	12.42	21.64	24.24	26.63	38.90	
blood_group	0	1	13.80	1.84	11.00	13.00	14.00	15.00	18.00	
pulse_rate_bpm	0	1	73.25	4.43	13.00	72.00	72.00	74.00	82.00	
rr_breaths_min	0	1	19.24	1.69	16.00	18.00	18.00	20.00	28.00	
hb_g_dl	0	1	11.16	0.87	8.50	10.50	11.00	11.70	14.80	
cycle_r_i	0	1	2.56	0.90	2.00	2.00	2.00	4.00	5.00	
cycle_length_days	0	1	4.94	1.49	0.00	4.00	5.00	5.00	12.00	
marraige_status_yrs	1	1	7.68	4.80	0.00	4.00	7.00	10.00	30.00	
pregnant_y_n	0	1	0.38	0.49	0.00	0.00	0.00	1.00	1.00	
no_of_aborptions	0	1	0.29	0.69	0.00	0.00	0.00	0.00	5.00	
i_beta_hcg_m_iu_m_l	0	1	664.55	3348.92	1.30	1.99	20.00	297.21	32460.97	
fsh_m_iu_m_l	0	1	14.60	217.02	0.21	3.30	4.85	6.41	5052.00	
lh_m_iu_m_l	0	1	6.47	86.67	0.02	1.02	2.30	3.68	2018.00	
fsh_lh	0	1	6.90	60.69	0.00	1.42	2.17	3.96	1372.83	
hip_inch	0	1	37.99	3.97	26.00	36.00	38.00	40.00	48.00	
waist_inch	0	1	33.84	3.60	24.00	32.00	34.00	36.00	47.00	
waist_hip_ratio	0	1	0.89	0.05	0.76	0.86	0.89	0.93	0.98	
tsh_m_iu_l	0	1	2.98	3.76	0.04	1.48	2.26	3.57	65.00	
prl_ng_m_l	0	1	24.32	14.97	0.40	14.52	21.92	29.89	128.24	
vit_d3_ng_m_l	0	1	49.92	346.21	0.00	20.80	25.90	34.50	6014.66	
prg_ng_m_l	0	1	0.61	3.81	0.05	0.25	0.32	0.45	85.00	
rbs_mg_dl	0	1	99.84	18.56	60.00	92.00	100.00	107.00	350.00	
weight_gain_y_n	0	1	0.38	0.49	0.00	0.00	0.00	1.00	1.00	
hair_growth_y_n	0	1	0.27	0.45	0.00	0.00	0.00	1.00	1.00	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
skin_darkening_y_n	0	1	0.31	0.46	0.00	0.00	0.00	1.00	1.00	
hair_loss_y_n	0	1	0.45	0.50	0.00	0.00	0.00	1.00	1.00	
pimples_y_n	0	1	0.49	0.50	0.00	0.00	0.00	1.00	1.00	
fast_food_y_n	1	1	0.51	0.50	0.00	0.00	1.00	1.00	1.00	
reg_exercise_y_n	0	1	0.25	0.43	0.00	0.00	0.00	0.00	1.00	
bp_systolic_mm_hg	0	1	114.66	7.38	12.00	110.00	110.00	120.00	140.00	
bp_diastolic_mm_hg	0	1	76.93	5.57	8.00	70.00	80.00	80.00	100.00	
follicle_no_l	0	1	6.13	4.23	0.00	3.00	5.00	9.00	22.00	
follicle_no_r	0	1	6.64	4.44	0.00	3.00	6.00	10.00	20.00	
avg_f_size_l_mm	0	1	15.02	3.57	0.00	13.00	15.00	18.00	24.00	
avg_f_size_r_mm	0	1	15.45	3.32	0.00	13.00	16.00	18.00	24.00	
endometrium_mm	0	1	8.48	2.17	0.00	7.00	8.50	9.80	18.00	

The overview of the data reveals that sl_no and patient_file seem to have the same information. The majority of observations in the last column (x45) are missing (539 out of 541). The variables 'sl_no' and 'x45' have therefore been removed.

The names of the variables include the units of measure, making the variable names complex. These have been simplified to facilitate analysis. Finally, variables were mutated to the correct data types and factor levels have been specified.

```
# Confirm that sl_no and patient_file_no are the same column
all(identical(data$sl_no, data$patient_file_no))
```

```
## [1] TRUE
```

```
# Remove variables 'sl_no' and 'x45'
data <- subset(data, select = -c(sl_no,x45))

# Rename variables to simplify them
data <- data %>%
  rename(
    id = patient_file_no,
    pcos = pcos_y_n,
    age = age_yrs,
    weight = weight_kg,
    height = height_cm,
    pulse_rate = pulse_rate_bpm,
    rr = rr_breaths_min,
    hb = hb_g_dl,
    cycle = cycle_r_i,
    cycle_length = cycle_length_days,
    marriage_status = marraige_status_yrs,
    pregnant = pregnant_y_n,
    no_of_abortions = no_of_aborptions,
    i_betahcg = i_beta_hcg_m_iu_m_l,
    ii_betahcg = ii_beta_hcg_m_iu_m_l,
    fsh = fsh_m_iu_m_l,
    lh =lh_m_iu_m_l,
    fsh_lh_ratio = fsh_lh,
    hip = hip_inch,
    waist = waist_inch,
    tsh = tsh_m_iu_l,
    amh = amh_ng_m_l,
    prl = prl_ng_m_l,
    vitd3 = vit_d3_ng_m_l,
    prg = prg_ng_m_l,
    rbs = rbs_mg_dl,
    weight_gain = weight_gain_y_n,
    hair_growth = hair_growth_y_n,
    skin_darkening = skin_darkening_y_n,
    hair_loss = hair_loss_y_n,
    pimples = pimples_y_n,
    fast_food = fast_food_y_n,
    reg_exercise = reg_exercise_y_n,
    bp_systolic = bp_systolic_mm_hg,
    bp_diastolic = bp_diastolic_mm_hg,
    avg_f_size_l = avg_f_size_l_mm,
    avg_f_size_r = avg_f_size_r_mm,
    endometrium = endometrium_mm
    )

# Mutate variables incorrectly labelled as character to numeric and those
# incorrectly labelled as numeric to factors
data = data %>%
  mutate(id = as.character(id),
         pcos = as.factor(pcos),
         ii_betahcg = case_when(ii_betahcg == "1.99."~ "1.99", #I found
                                          #this typo when exploring the missing data and
                                          # checking the excel file of said individual
                                          TRUE ~ ii_betahcg),
         ii_betahcg = as.numeric(ii_betahcg),
         amh = as.numeric(amh),
         blood_group = as.factor (blood_group),
         pregnant = as.factor(pregnant),
         weight_gain = as.factor(weight_gain),
         hair_growth = as.factor(hair_growth),
         skin_darkening = as.factor(skin_darkening),
         hair_loss = as.factor(hair_loss),
         pimples = as.factor(pimples),
         fast_food = as.factor(fast_food),
         reg_exercise = as.factor(reg_exercise))
```

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

```
# The levels of binary variables are set to 'No' and 'Yes' to assist analysis later
levels(data$pcos)=c("No","Yes")
levels(data$pregnant)=c("No","Yes")
levels(data$weight_gain)=c("No","Yes")
levels(data$hair_growth)=c("No","Yes")
levels(data$skin_darkening)=c("No","Yes")
levels(data$hair_loss)=c("No","Yes")
levels(data$pimples)=c("No","Yes")
levels(data$fast_food)=c("No","Yes")
```

```r
levels(data$reg_exercise)=c("No","Yes")
levels(data$blood_group)=c("A+","A-","B+","B-","O+","O-","AB+","AB-")

# Overview of the cleaned data
skim(data)
```

Table 4: Data summary

Name	data
Number of rows	541
Number of columns	43
Column type frequency:	
character	1
factor	10
numeric	32
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
id	0	1	1	3	0	541	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
pcos	0	1	FALSE	2	No: 364, Yes: 177
blood_group	0	1	FALSE	8	O+: 206, B+: 135, A+: 108, AB+: 42
pregnant	0	1	FALSE	2	No: 335, Yes: 206
weight_gain	0	1	FALSE	2	No: 337, Yes: 204
hair_growth	0	1	FALSE	2	No: 393, Yes: 148
skin_darkening	0	1	FALSE	2	No: 375, Yes: 166
hair_loss	0	1	FALSE	2	No: 296, Yes: 245
pimples	0	1	FALSE	2	No: 276, Yes: 265
fast_food	1	1	FALSE	2	Yes: 278, No: 262
reg_exercise	0	1	FALSE	2	No: 407, Yes: 134

Variable type: numeric

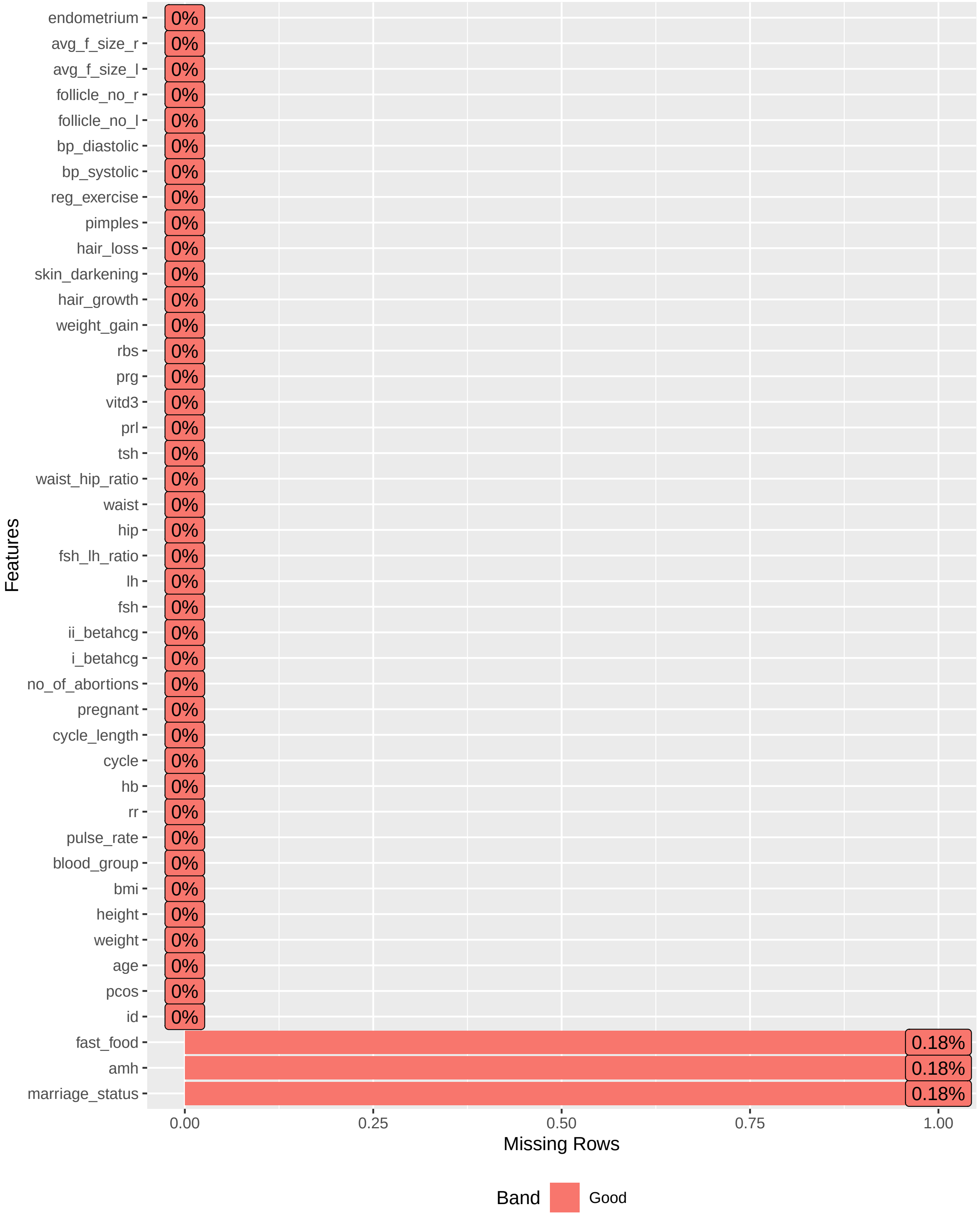
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	31.43	5.41	20.00	28.00	31.00	35.00	48.00	
weight	0	1	59.64	11.03	31.00	52.00	59.00	65.00	108.00	
height	0	1	156.48	6.03	137.00	152.00	156.00	160.00	180.00	
bmi	0	1	24.31	4.06	12.42	21.64	24.24	26.63	38.90	
pulse_rate	0	1	73.25	4.43	13.00	72.00	72.00	74.00	82.00	
rr	0	1	19.24	1.69	16.00	18.00	18.00	20.00	28.00	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
hb	0	1	11.16	0.87	8.50	10.50	11.00	11.70	14.80	
cycle	0	1	2.56	0.90	2.00	2.00	2.00	4.00	5.00	
cycle_length	0	1	4.94	1.49	0.00	4.00	5.00	5.00	12.00	
marriage_status	1	1	7.68	4.80	0.00	4.00	7.00	10.00	30.00	
no_of_abortions	0	1	0.29	0.69	0.00	0.00	0.00	0.00	5.00	
i_betahcg	0	1	664.55	3348.92	1.30	1.99	20.00	297.21	32460.97	
ii_betahcg	0	1	238.23	1603.83	0.99	1.99	1.99	97.63	25000.00	
fsh	0	1	14.60	217.02	0.21	3.30	4.85	6.41	5052.00	
lh	0	1	6.47	86.67	0.02	1.02	2.30	3.68	2018.00	
fsh_lh_ratio	0	1	6.90	60.69	0.00	1.42	2.17	3.96	1372.83	
hip	0	1	37.99	3.97	26.00	36.00	38.00	40.00	48.00	
waist	0	1	33.84	3.60	24.00	32.00	34.00	36.00	47.00	
waist_hip_ratio	0	1	0.89	0.05	0.76	0.86	0.89	0.93	0.98	
tsh	0	1	2.98	3.76	0.04	1.48	2.26	3.57	65.00	
amh	1	1	5.62	5.88	0.10	2.01	3.70	6.93	66.00	
prl	0	1	24.32	14.97	0.40	14.52	21.92	29.89	128.24	
vitd3	0	1	49.92	346.21	0.00	20.80	25.90	34.50	6014.66	
prg	0	1	0.61	3.81	0.05	0.25	0.32	0.45	85.00	
rbs	0	1	99.84	18.56	60.00	92.00	100.00	107.00	350.00	
bp_systolic	0	1	114.66	7.38	12.00	110.00	110.00	120.00	140.00	
bp_diastolic	0	1	76.93	5.57	8.00	70.00	80.00	80.00	100.00	
follicle_no_l	0	1	6.13	4.23	0.00	3.00	5.00	9.00	22.00	
follicle_no_r	0	1	6.64	4.44	0.00	3.00	6.00	10.00	20.00	
avg_f_size_l	0	1	15.02	3.57	0.00	13.00	15.00	18.00	24.00	
avg_f_size_r	0	1	15.45	3.32	0.00	13.00	16.00	18.00	24.00	
endometrium	0	1	8.48	2.17	0.00	7.00	8.50	9.80	18.00	

2.3 Missing values

The missing variables are checked using the plot below.

```
plot_missing(data)
```



One missing value is recorded for the variables fast_food, marriage_status and amh. The code below aims to determine whether the missing data points are all for the same person.

```
data %>%
  filter(is.na(fast_food) |
          is.na(marriage_status) |
          is.na(amh)) %>%
  dplyr::select(c(id, fast_food, marriage_status, amh)) %>%
  knitr::kable()
```

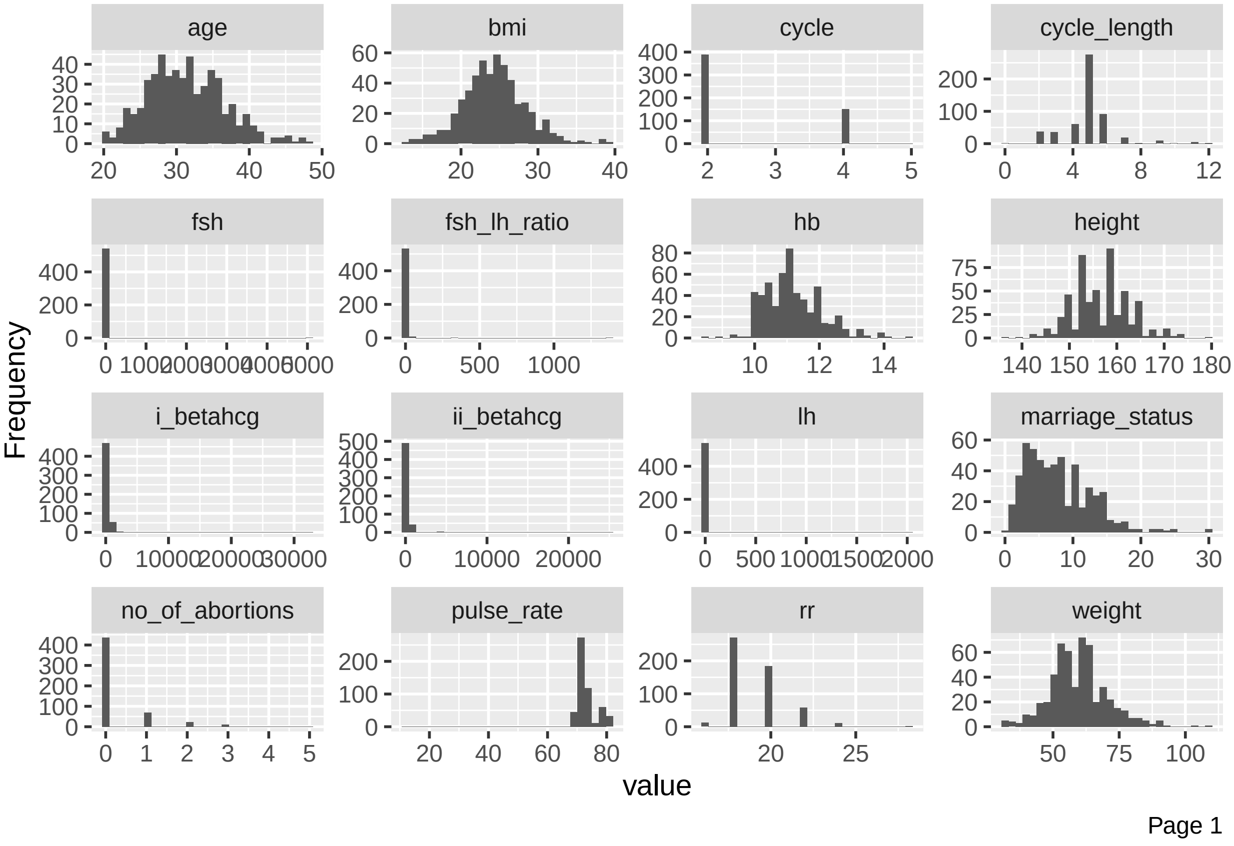
id	fast_food	marriage_status	amh
157	NA	5	5.27
306	No	9	NA
459	No	NA	6.60

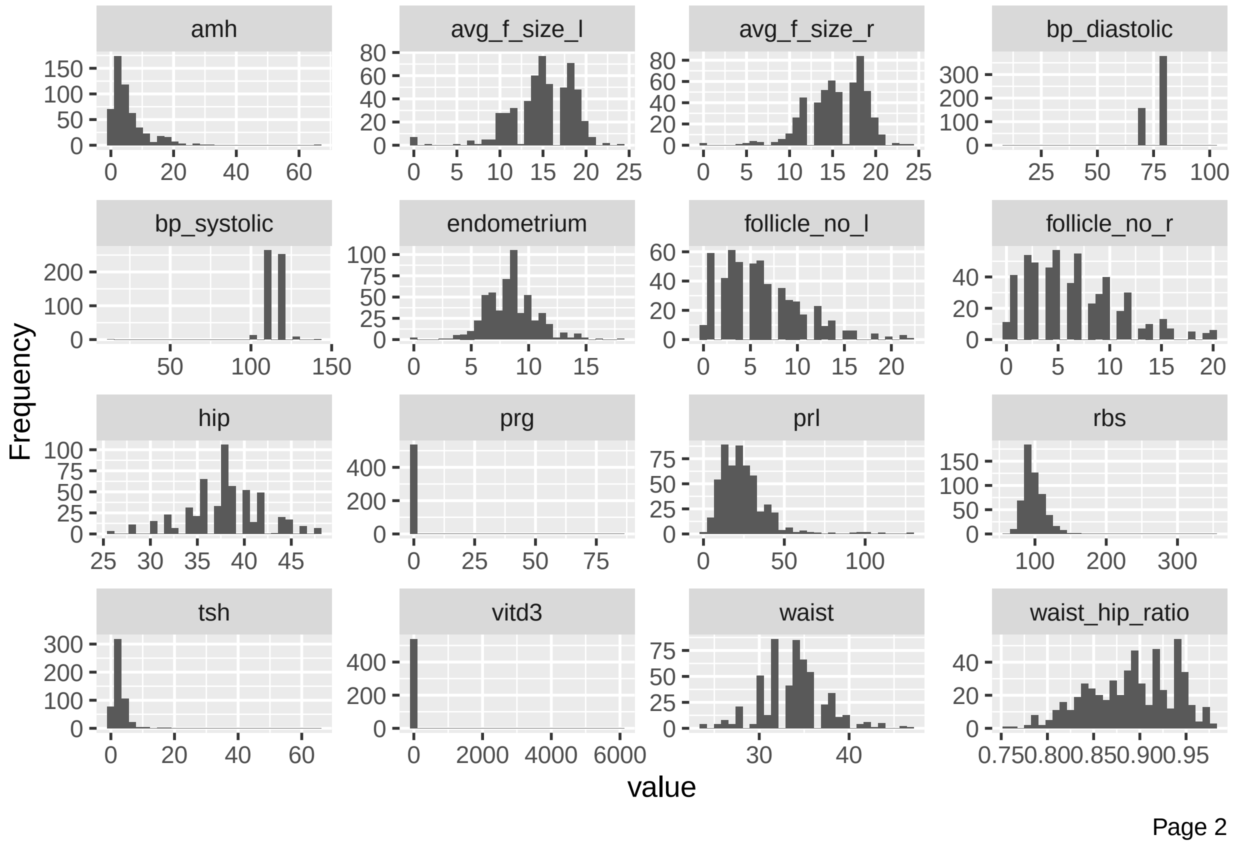
We can observe that the individuals with missing observations are different, and thus these observations (rows) do not need to be removed at this stage. If missing variables end up in the training data set, these will be managed with multiple imputation.

3 Variation in continuous variables

The variation of continuous variables in the dataset are explored using the histograms below.

```
plot_histogram(data)
```





From the plots, we can appreciate a few key observations. The patients are aged between 20 and 48 with a large proportion being classed according to their body mass index (BMI) as having a healthy weight (18.5-24.9) or being overweight (25.0-29.9). Most women are either not married or have been married for less than 10 years, and the majority have not had a miscarriage/abortion. Many continuous variables seem to follow a normal distribution, however, several variables seem to suffer from little to no variation, possibly due to outliers. Outliers will be examined in detail.

3.1 Outliers

We can observe that the variables prg, vit_d3, fsh_lh, fsh, i_betahcg, ii_betahcg, lh and pulse_rate seem to have no variation. This could be happening because of the presence of outliers that make the data look like if it were invariant, or because of the data is not normally distributed in these variables. This can be checked by observing the summary of said variables. These are explored in detail.

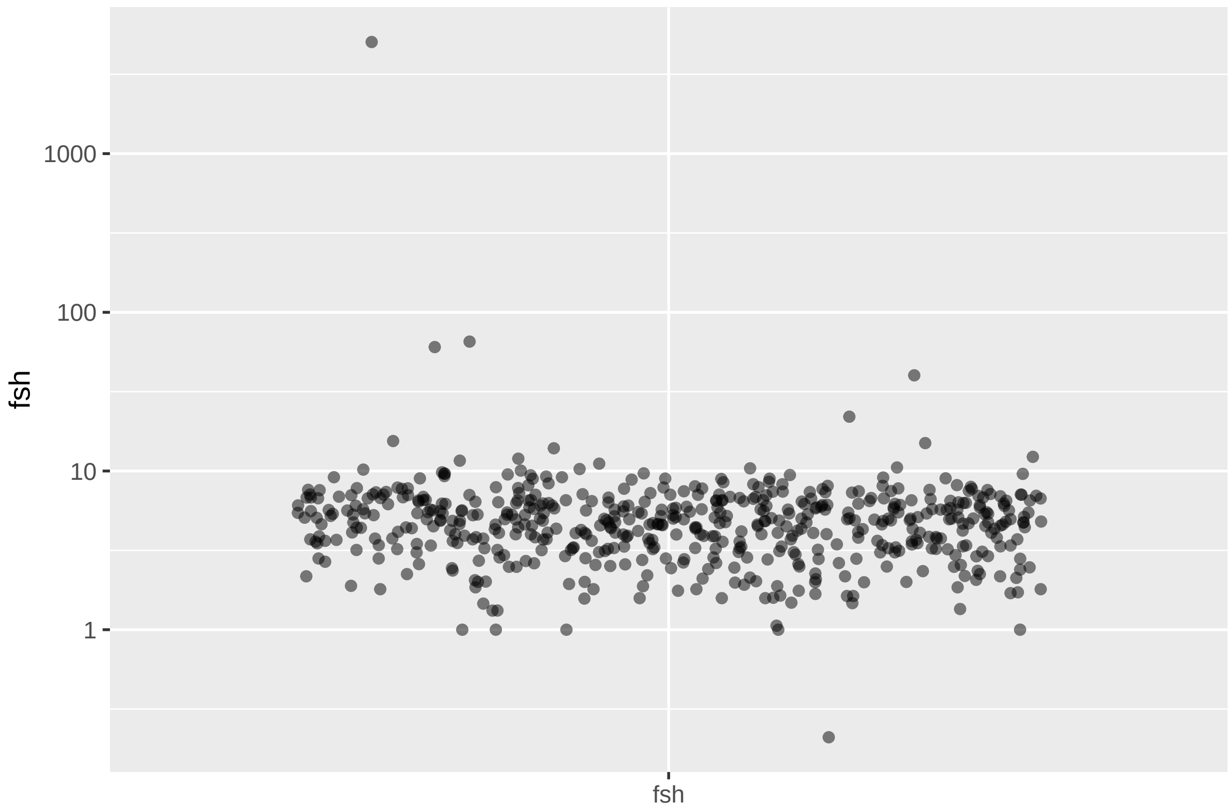
```r
data %>%
  dplyr::select(prg, vitd3, fsh_lh_ratio, fsh, i_betahcg, ii_betahcg, lh, pulse_rate) %>%
  rownames_to_column(var = "ID") %>%
  pivot_longer(-ID, names_to = "variables", values_to = "data") %>%
  group_by(variables) %>%
  summarise(mean = mean(data, na.rm = TRUE),
            q1 = quantile(data, 0.25),
            median = quantile(data, 0.5),
            q3 = quantile(data,0.75),
            max = max(data),
            min = min(data)) %>%
  knitr::kable()
```

variables	mean	q1	median	q3	max	min
fsh	14.6018318	3.300000	4.850000	6.410000	5052.000	0.2100000
fsh_lh_ratio	6.9048308	1.416244	2.169231	3.959184	1372.826	0.0021457
i_betahcg	664.5492348	1.990000	20.000000	297.210000	32460.970	1.3000000
ii_betahcg	238.2329926	1.990000	1.990000	97.630000	25000.000	0.9900000
lh	6.4699187	1.020000	2.300000	3.680000	2018.000	0.0200000
prg	0.6109445	0.250000	0.320000	0.450000	85.000	0.0470000
pulse_rate	73.2476895	72.000000	72.000000	74.000000	82.000	13.0000000
vitd3	49.9158743	20.800000	25.900000	34.500000	6014.660	0.0000000

By looking at the median and quartiles, we can observe that the data looks not to be normally distributed because there are outliers that drag the distributions. I will check which samples are outliers for each of these variables.

3.1.1. FSH hormone Now, I will look for outliers in the FSH hormone

```
## FSH hormone
data %>%
  ggplot(aes(x = "fsh", y = fsh)) +
  geom_jitter(alpha = 0.5) +
  scale_y_log10()+
  xlab("")
```



```
data %>%
  filter(fsh > 1000) %>%
  pull(id)
```

```
## [1] "330"
```

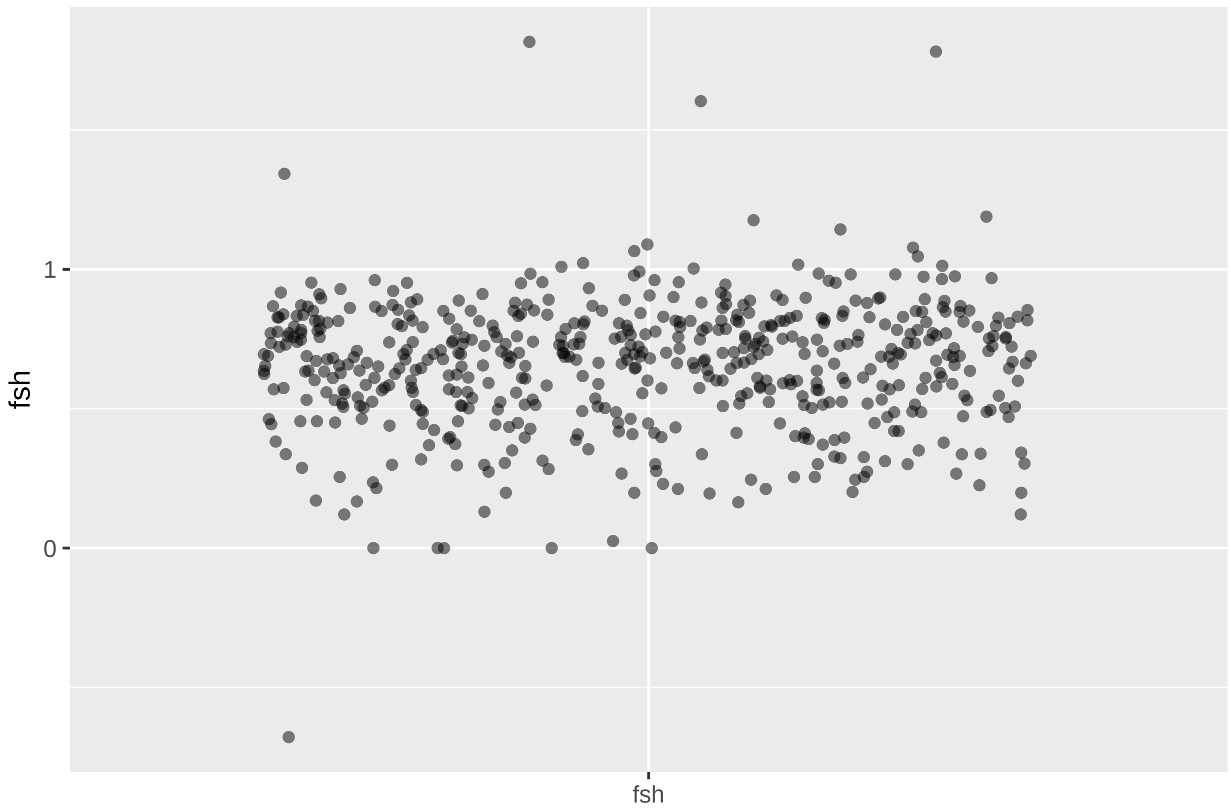
According to reference values, this sample has an impossible biological value. Therefore, this observation and related variables will be set to NA.

```
data[data$id == 330,"fsh" ] = NA
data[data$id == 330,"fsh_lh_ratio" ] = NA
```

The values will be log-10 transformed because looking at the quartiles above, data is compressed in the left side of the histogram

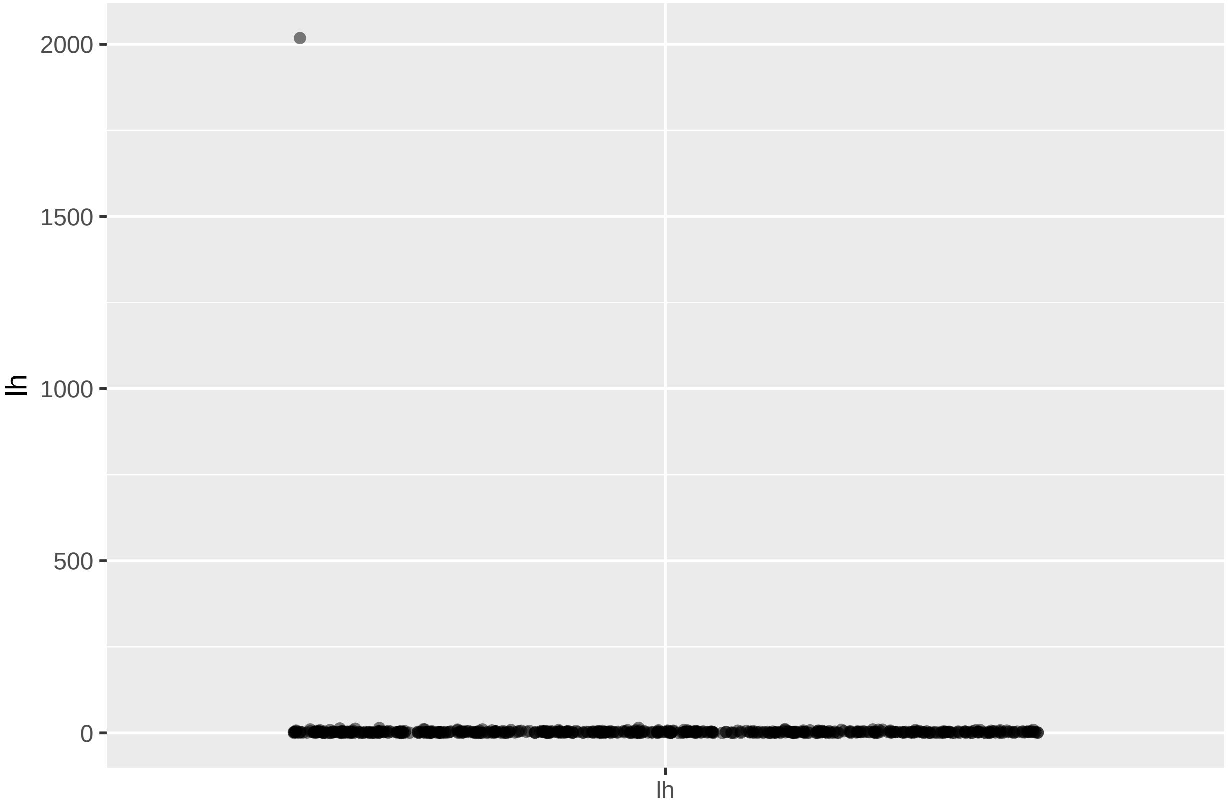
```
data = data %>%
  mutate(fsh = log10(fsh))

data %>%
  ggplot(aes(x = "fsh", y = fsh)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



3.1.2. LH hormone Now, I will look for outliers in the LH hormone

```
## LH hormone
data %>%
  ggplot(aes(x = "lh", y = lh)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```



```
data %>%
  filter(lh > 1000) %>%
  pull(id)
```

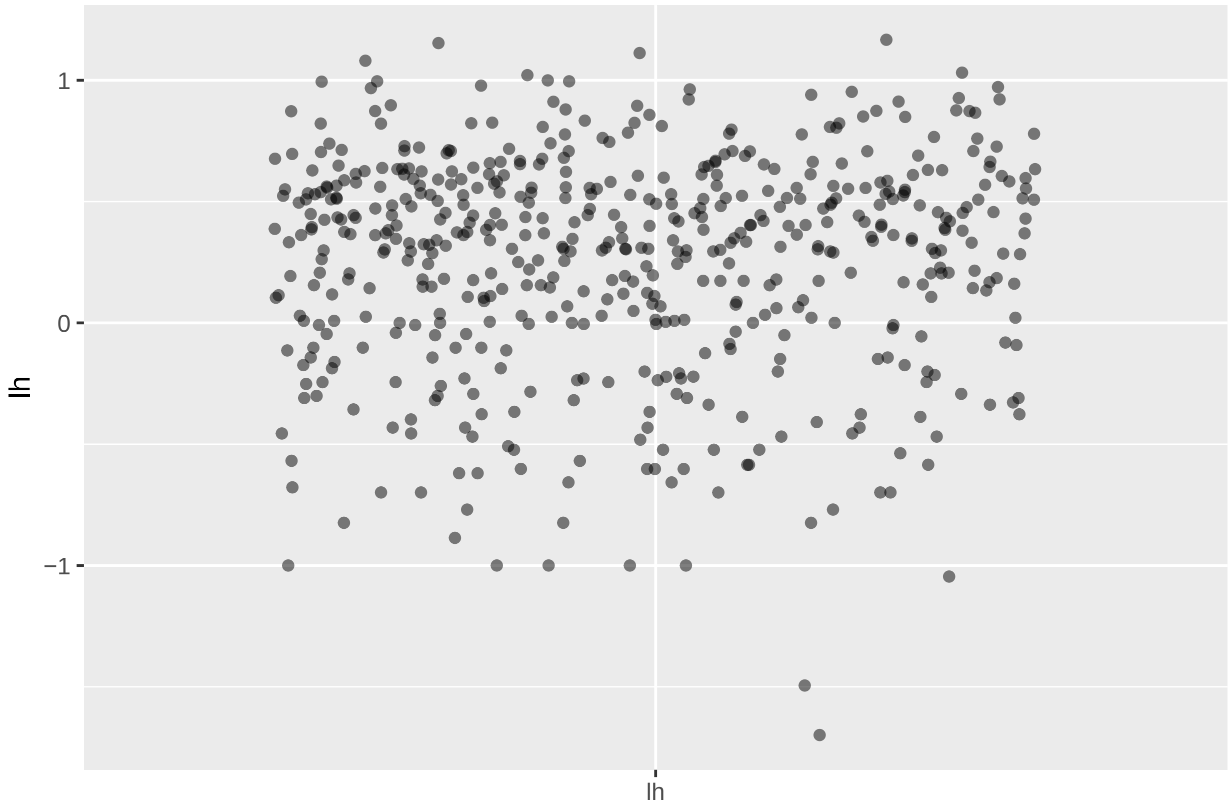
```
## [1] "456"
```

The individual 456 has a LH level outside of the reported reference levels. Therefore, this value will be set to NA and the variable will be log10-transformed. It is worth noticing that this individual is different to the one that had an anomalous FSH level, which supports the hypothesis if these values being technical mistakes.

```
data[data$id == 456,"lh" ] = NA
data[data$id == 456,"fsh_lh_ratio" ] = NA

data = data %>%
  mutate(lh = log10(lh))
```

```
data %>%
  ggplot(aes(x = "lh", y = lh)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```

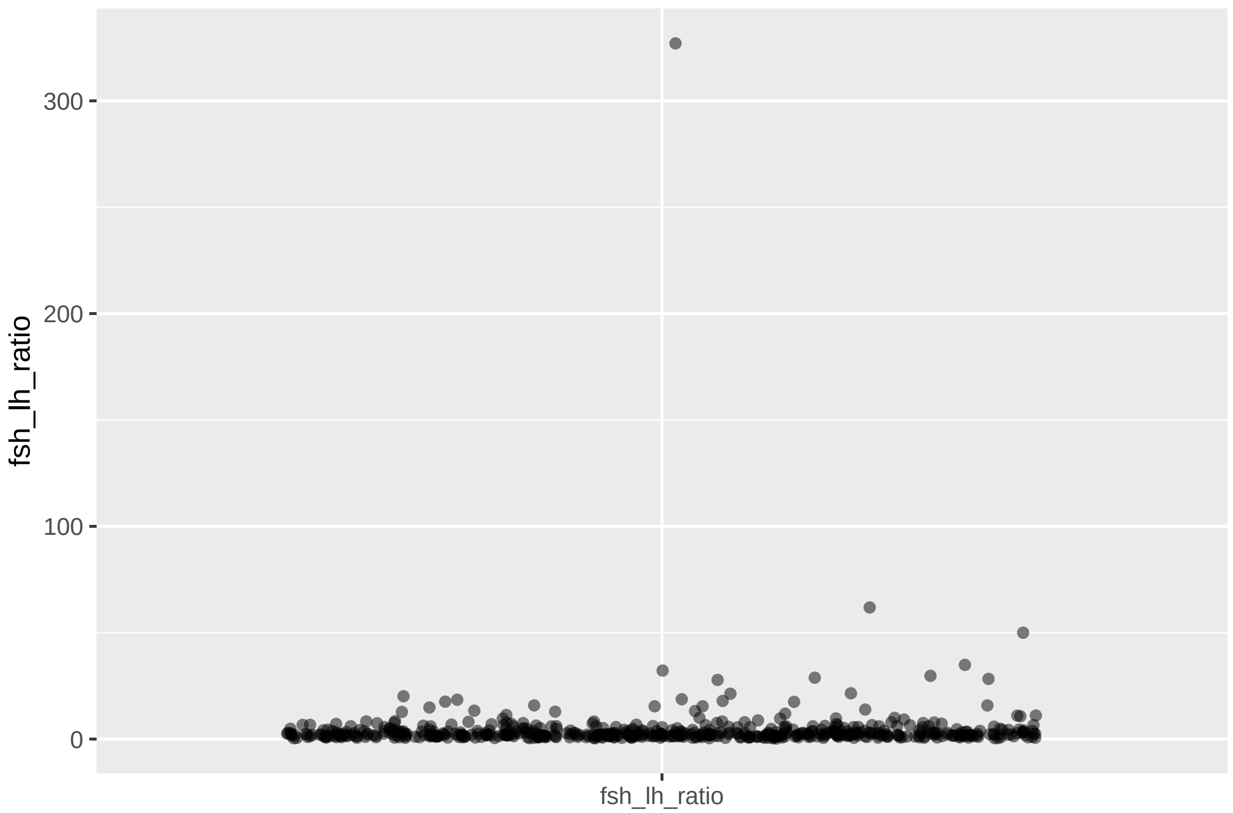
```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



3.1.3. FSH/LH ratio Since I have already removed outliers from the FSH and LH variables, the remaining outlier here should be occurring biologically. Therefore, this variable has simply been log-10 transformed.

```
## FSH/LH ratio
data %>%
  ggplot(aes(x = "fsh_lh_ratio", y = fsh_lh_ratio)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```

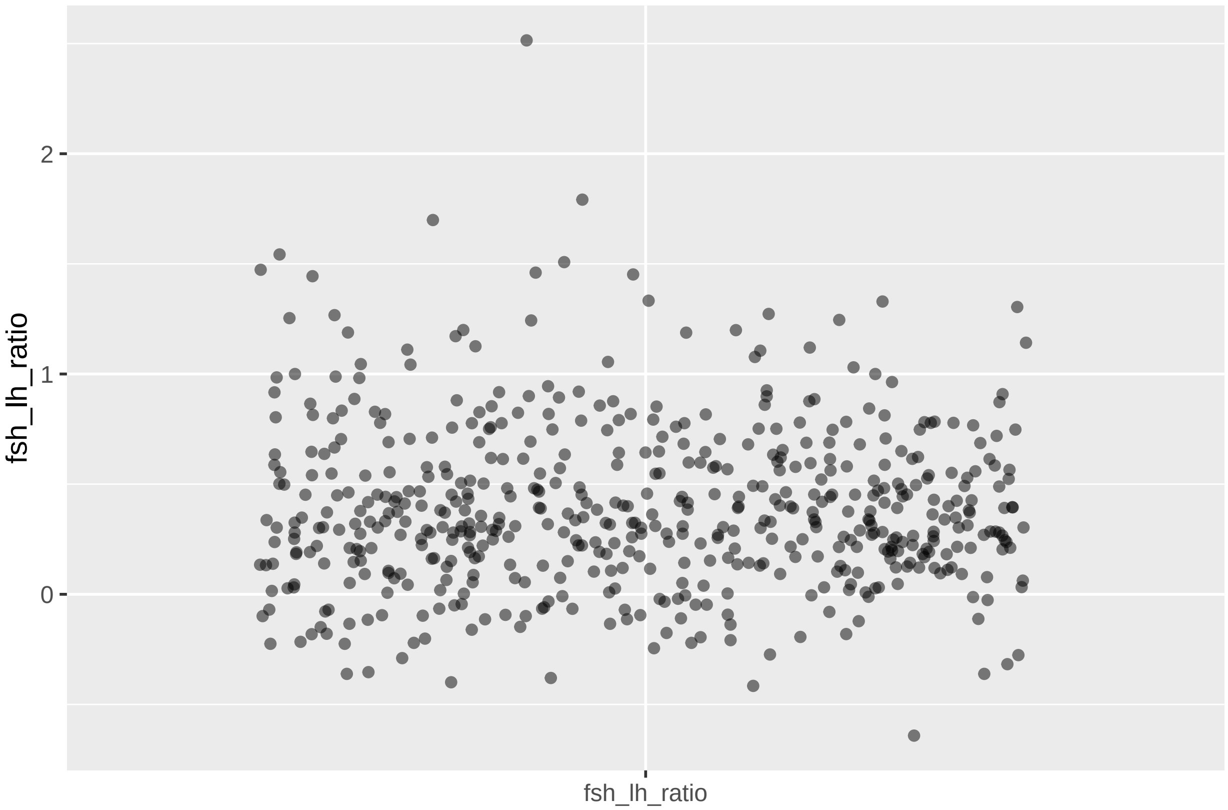


```
data %>%
  filter(fsh_lh_ratio > 250) %>%
  pull(id)
```

```
## [1] "251"
```

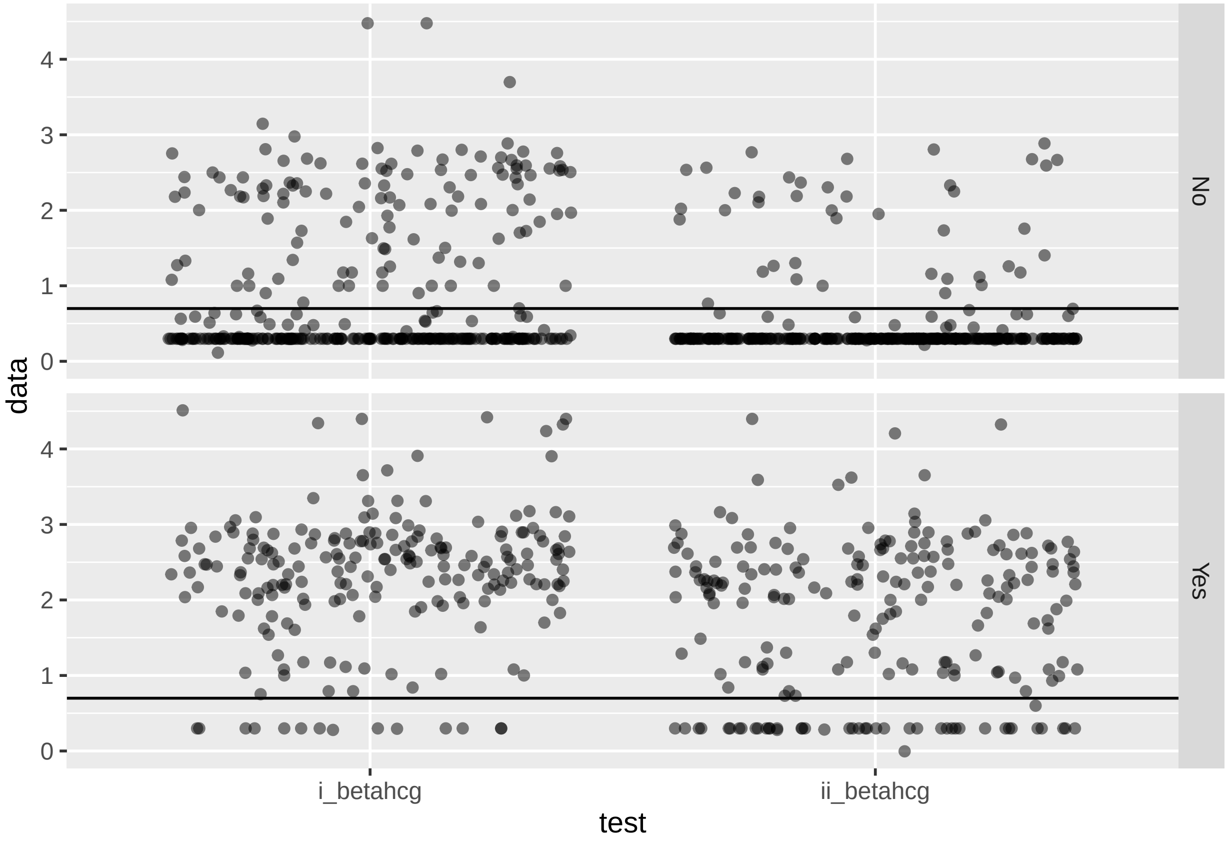
```
#I will flag this patient in case it pops out somewhere else in the analysis.

data = data %>%
  mutate(fsh_lh_ratio = log10(fsh_lh_ratio))

data %>%
  ggplot(aes(x = "fsh_lh_ratio", y = fsh_lh_ratio)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```



3.1.4. Human chorionic gonadotropin (hCG) in the blood According to reference levels, the values present in our dataset are within the expected biological range. Therefore, no outlers have been removed.

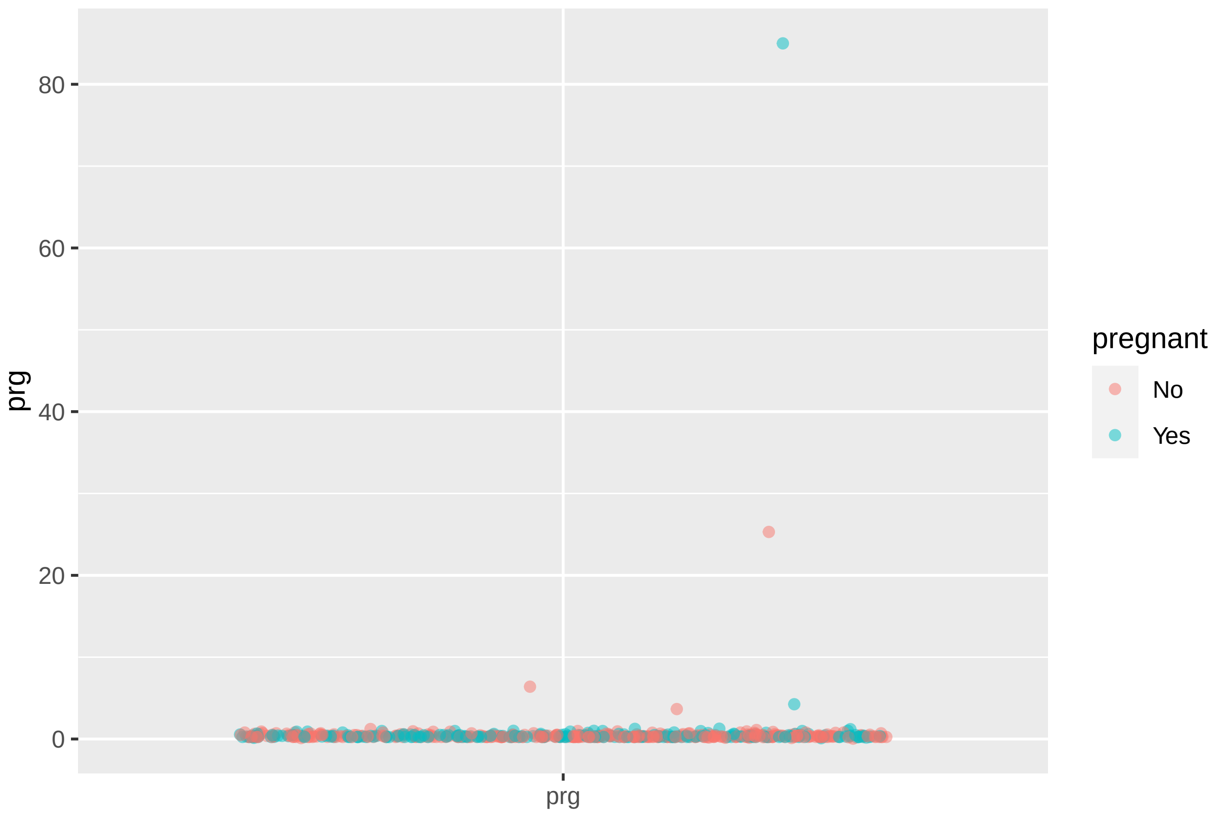
```
# Review the beta-HCG data (transformed for illustrative purposes) divided between the two tests and ba
data %>%
  dplyr::select(c(id, i_betahcg,ii_betahcg, pregnant)) %>%
  mutate(i_betahcg = log10(i_betahcg),
         ii_betahcg = log10(ii_betahcg)) %>%
  pivot_longer(- c(id,pregnant), names_to = "test", values_to = "data") %>%
  ggplot(aes(x = test, y = data)) +
  geom_jitter(alpha= 0.5) +
  geom_hline(yintercept = log10(5))+
  facet_grid("pregnant")
```



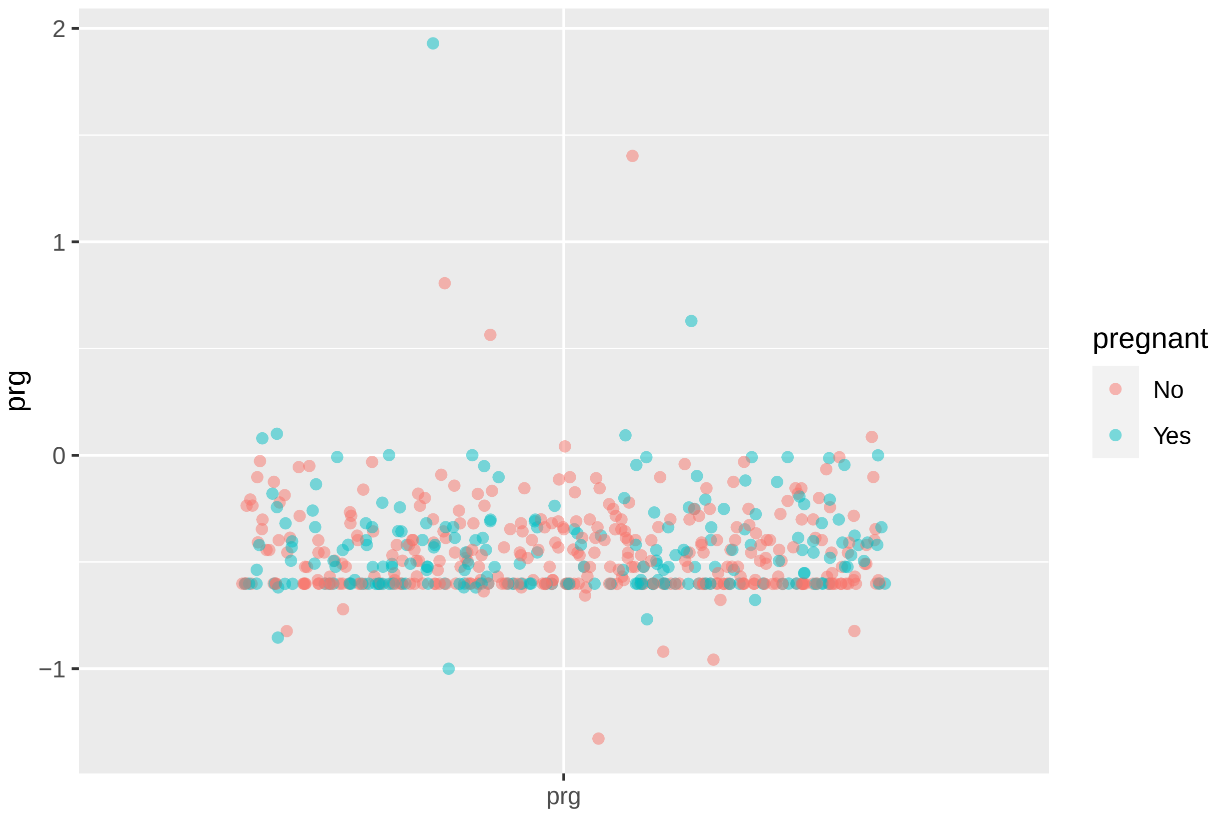
We can observe that even though both tests are supposed to measure the same hormone in blood, they do not provide similar results for many cases and there is no explanation in the data dictionary to indicate why two test results have been obtained - i.e. whether these are meaasured at different time points or whether different ways of testing beta-HCG were used. Due to this lack of information, using them in any models would be difficult as the interpretability and reproducibility of the model using these variables would be limited. It should be noted that non-pregnant women are supposed to have a beta-HCG level of less than 5 mIU/mL. However, this condition is not met for several non-pregnant women, even though overall levels in pregnant women seem to be higher. This may be because women did not know they were pregnant. Also, it is unclear from the data dictionary whether the pregnancy variable relates to women who are currently pregnant or have had previous pregnancies.

3.1.5. Progesterone According to reference levels, the values in the data set seem to be biologically possible. Therefore, only log-10 transformation of the data will be conducted.

```
data %>%
  ggplot(aes(x = "prg", y = prg, color = pregnant)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```



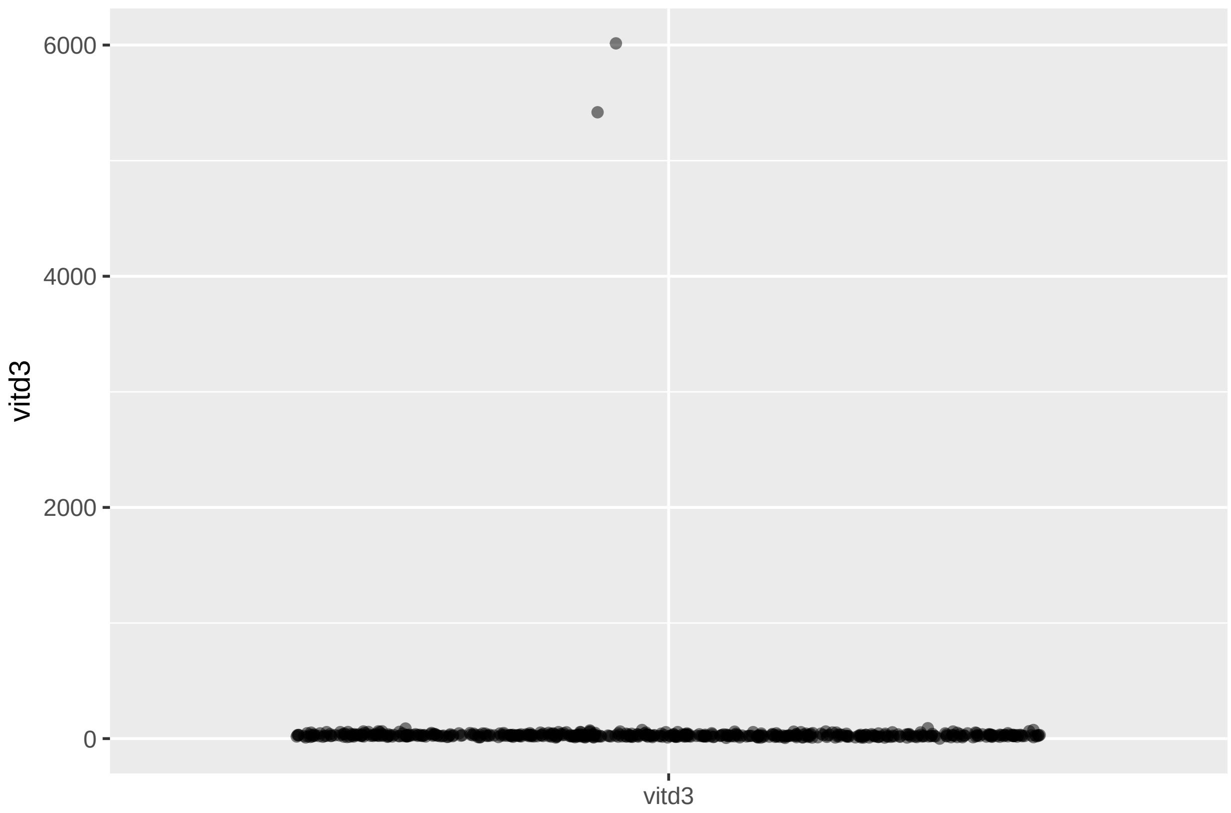
```
data = data %>%
  mutate(prg = log10(prg))

data %>%
  ggplot(aes(x = "prg", y = prg, color = pregnant)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```



Progesterone levels are very variable depending on the mentstrual cycle stage of the person at the time of the blood test, thus all values are plausible. Non-pregnant women might have a progesterone concentration of up to 25 ng/mL in the luteal stage of the menstrual cycle, which would explain the one high value in the plot above.

3.1.6. Vitamin D3 It has been reported that a normal range of vitamin D is 30 to 74 ng/mL, and that side effects and toxicity occur when blood concentrations reach 88 ng/mL or greater. Therefore, the outliers shown below with values over 5000 ng/mL are not biologically plausible and have therefore been set to NA.

```r
# Visualize values of Vitamin D3
data %>%
  ggplot(aes(x = "vitd3", y = vitd3)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```

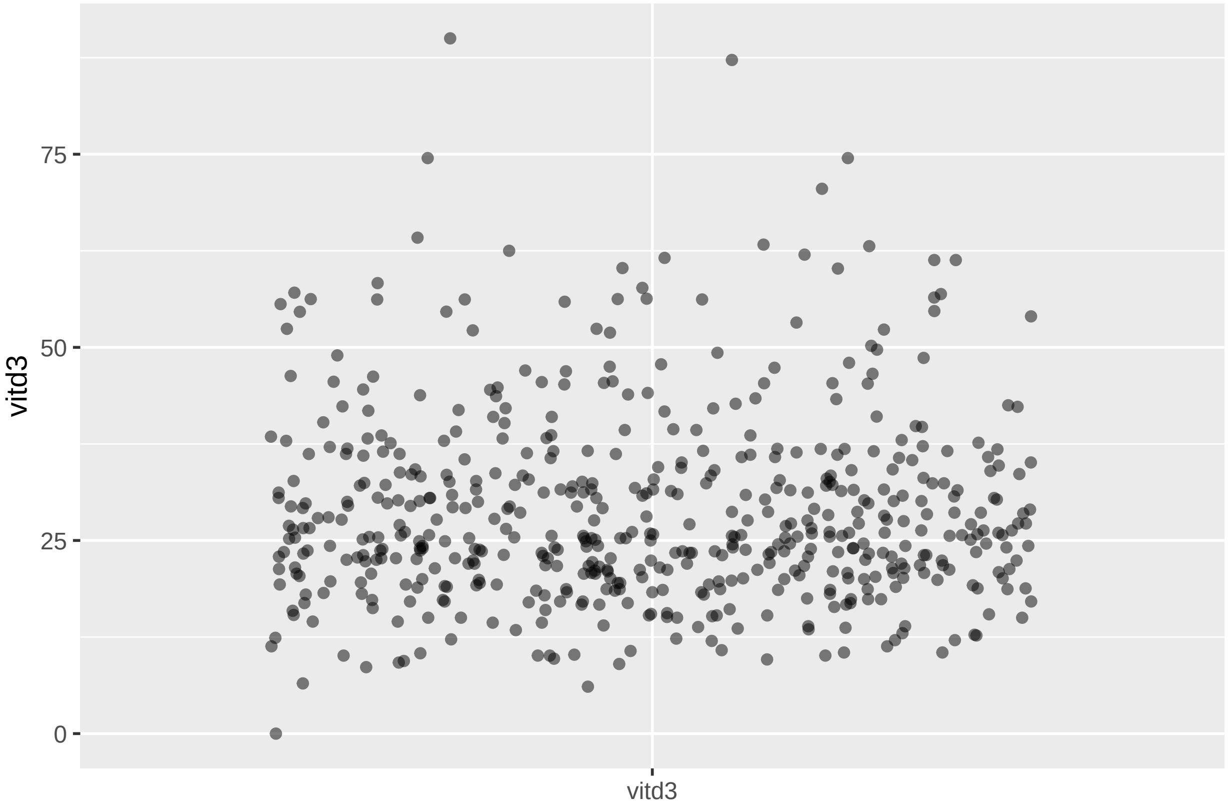


```
# Remove the biologically implausible values
data %>%
  filter(vitd3 > 90) %>%
  pull(id)
```

```
## [1] "192" "196"
```

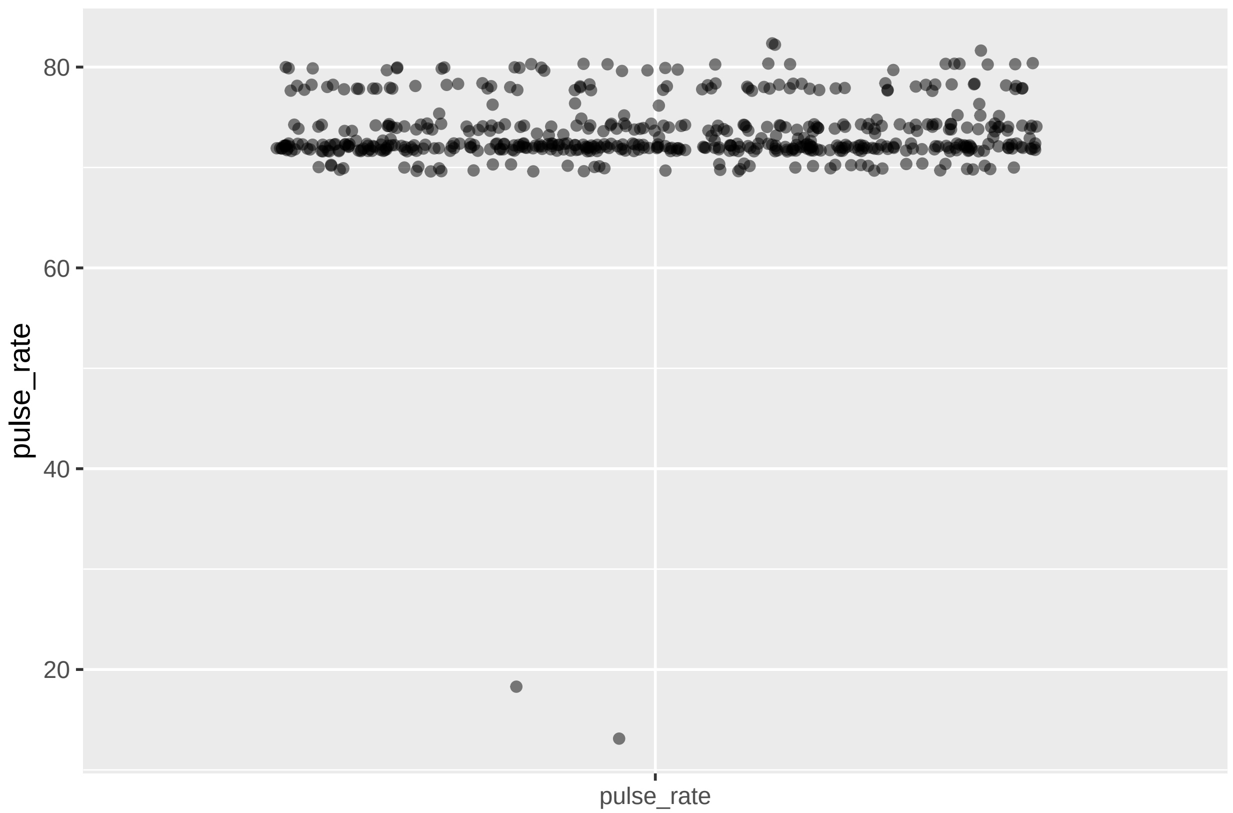
```
data[data$vitd3>90,"vitd3"] = NA

#Plot the data again

data %>%
  ggplot(aes(x = "vitd3", y = vitd3)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```



3.1.7. Pulse rate It is reported that the normal pulse rate goes from 60 to 100 bpm. Some atheletes can have a presting heart rate closer to 40, however, anything less is not compatible with life. Therefore, the two values below 20 found in our data set will be set to NA.

```
data %>%
  ggplot(aes(x = "pulse_rate", y = pulse_rate)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```

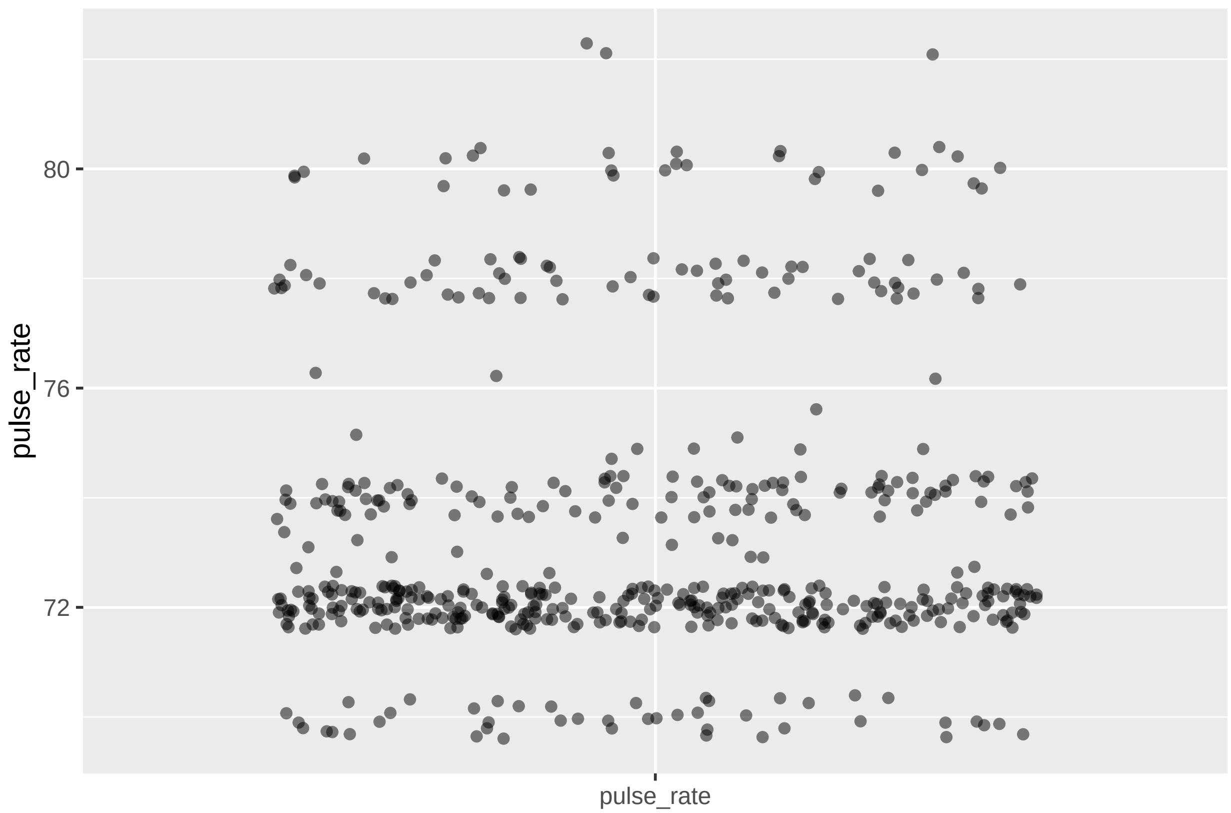


```
data %>%
  filter(pulse_rate < 40) %>%
  pull(id)
```

```
## [1] "224" "297"
```

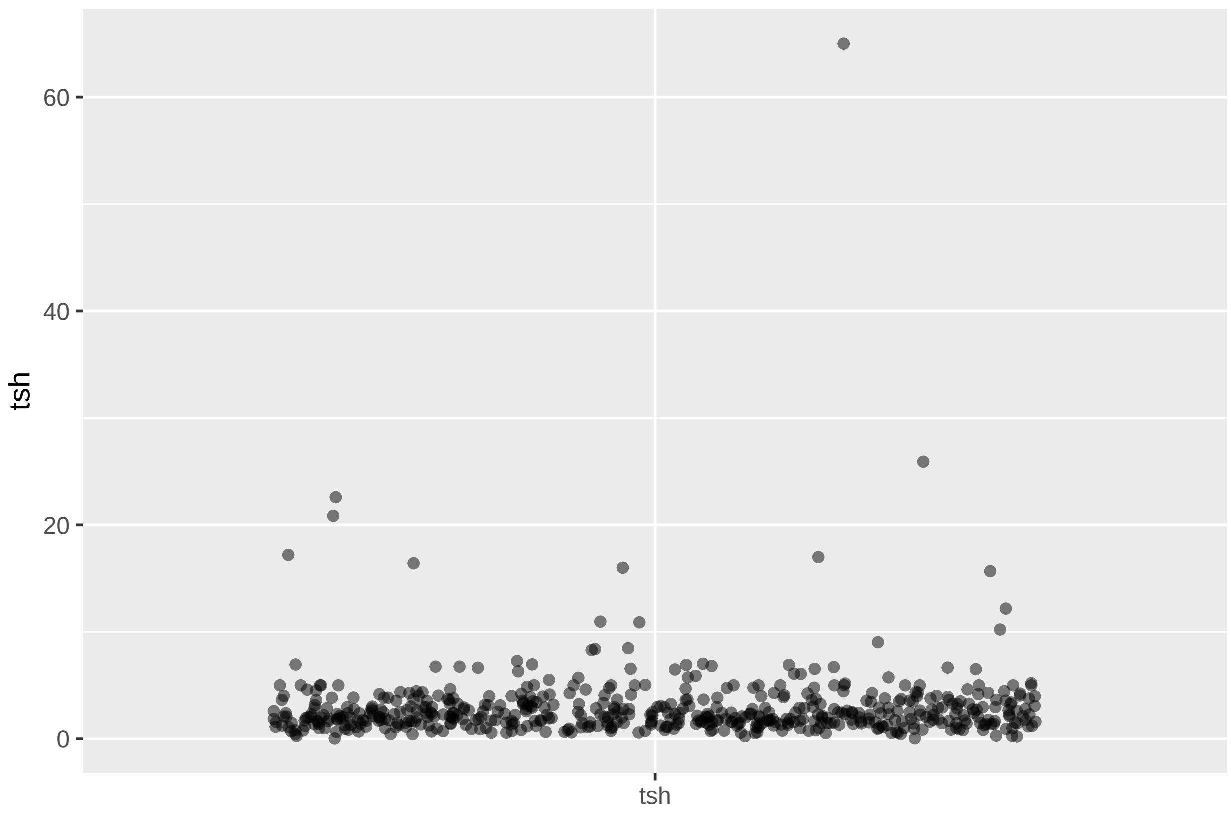
```
data[data$pulse_rate < 40,"pulse_rate"] = NA

#Re plot the data
data %>%
  ggplot(aes(x = "pulse_rate", y = pulse_rate)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```



3.1.8.Thyroid Stimulating Hormone (TSH) Now, I will look for outliers in the TSH hormone

```
data %>%
  ggplot(aes(x = "tsh", y = tsh)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```

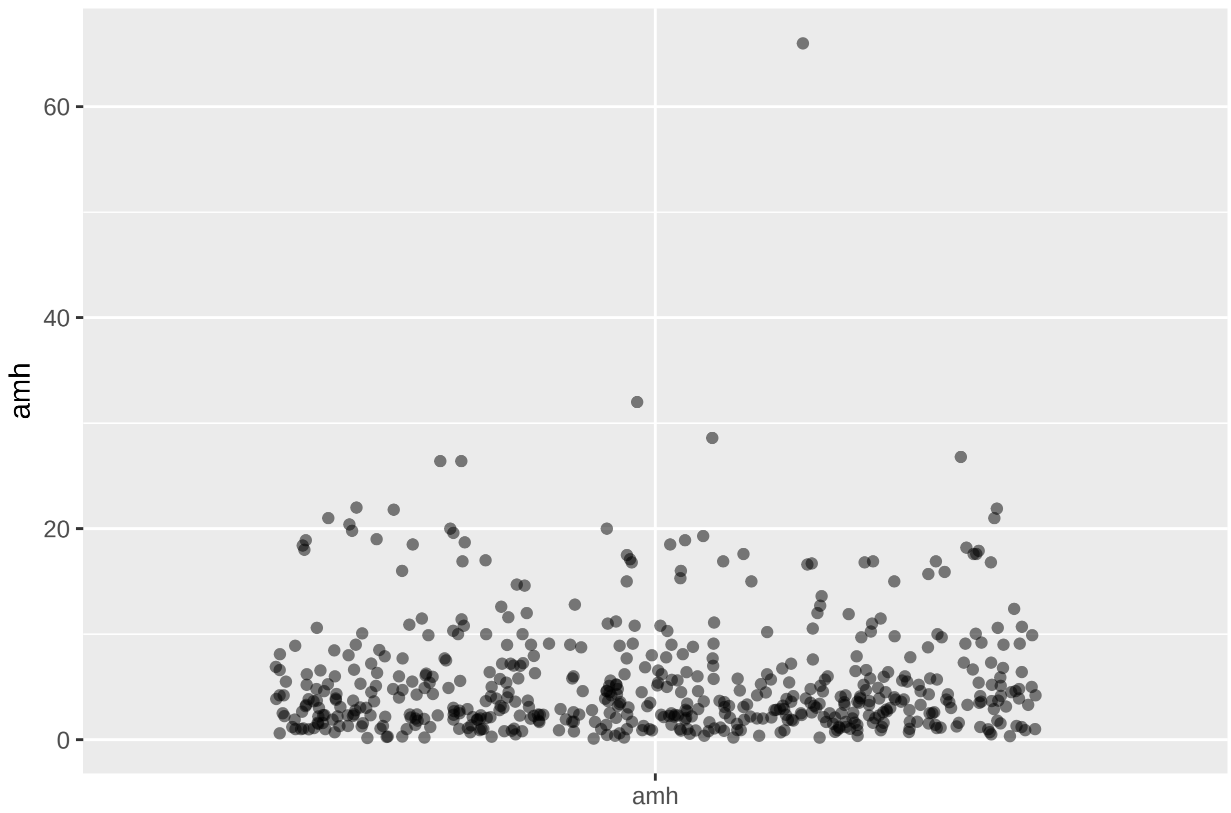


There is an outlier within our data, with a value of over 60 mIU/ml that is above the usual level in women with PCOS, which is around 6.4 ±4.2 mIU/L. However, values above 100 are encountered in clinical practice. Thus, this value seems biologically plausible and will be retained.

3.1.9. Anti-Mullerian Hormone (AMH) Now, I will look for outliers in the AMH hormone

```
data %>%
  ggplot(aes(x = "amh", y = amh)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



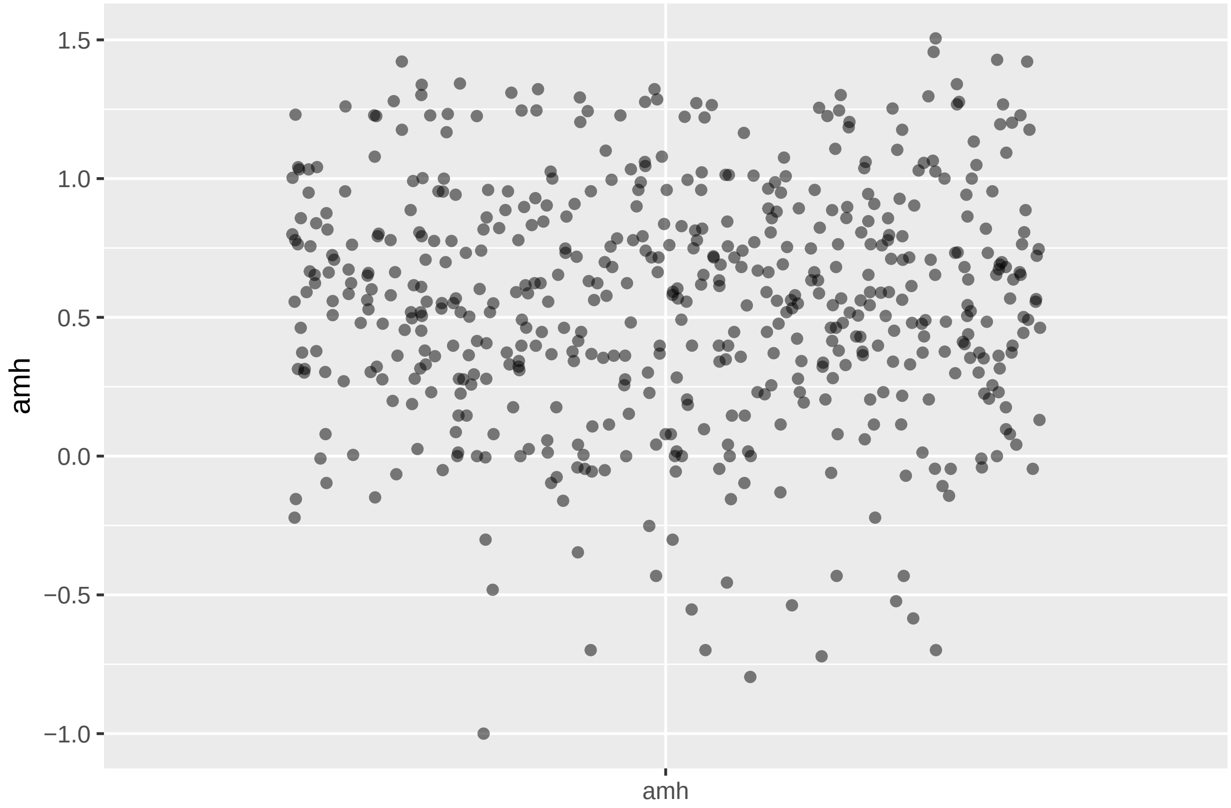
```
# Identify the patient with a high AMH level
data %>%
  filter(amh > 48) %>%
  pull(id)
```

```
## [1] "268"
```

I will remove the observation with AMH levels > 60 ng/mL since the reported values for women with PCOS have been reported to be around 4.32 ng/mL (2.633–7.777) in previous studies and even in studies of women with ultra high AMH values, the highest recorded value was 48 ng/ml.There are other values that seem to be too high, but only one is implausible and will therefore be set to NA.

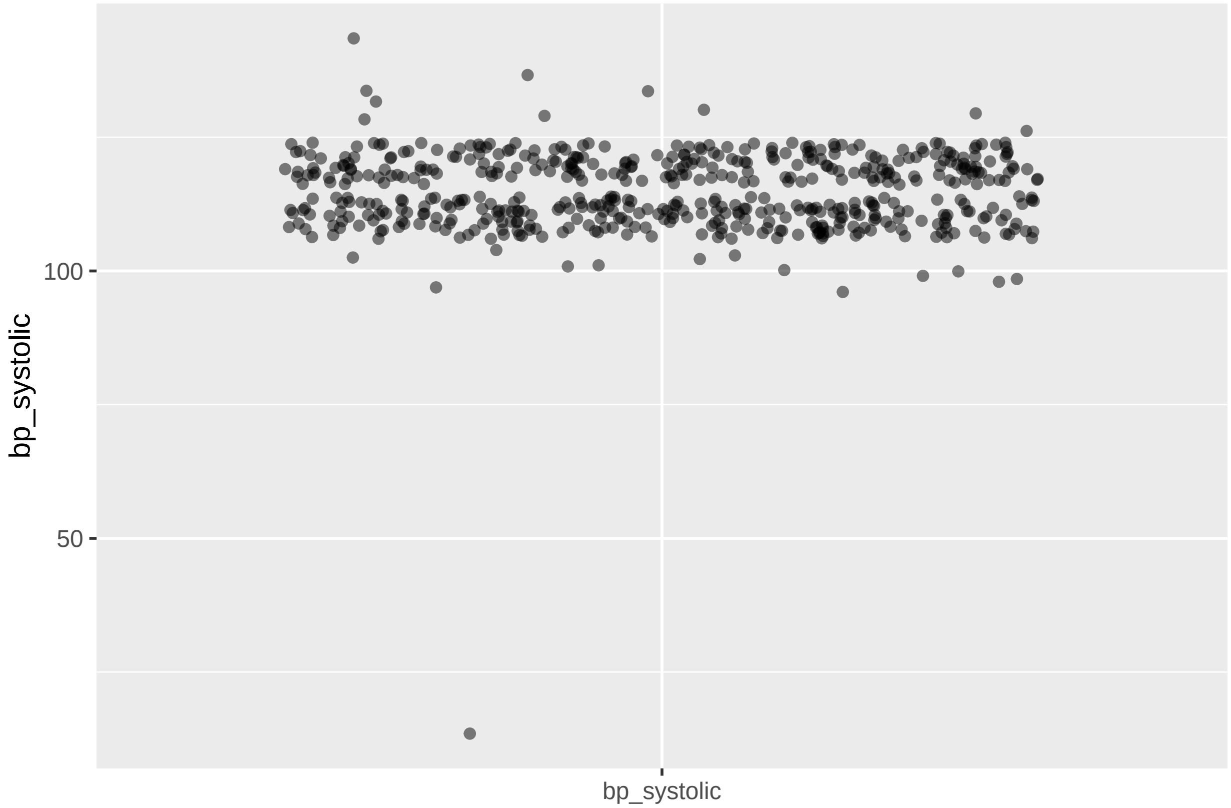
```
data[data$id == 268,"amh" ] = NA

#Transform the variable
data = data %>%
  mutate(amh = log10(amh))

#Re plot the data
data %>%
  ggplot(aes(x = "amh", y = amh)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```

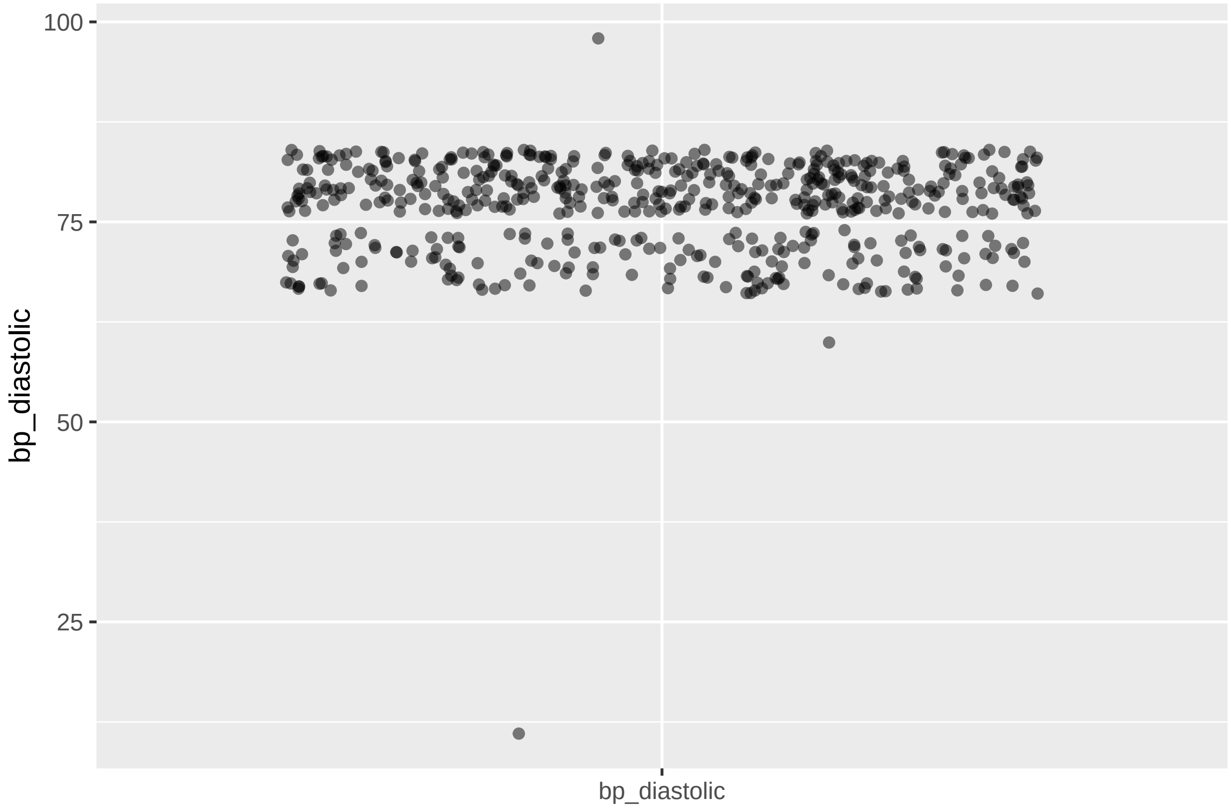


3.1.10. Blood pressure Now, I will look for outliers in the blood presure.

```
data %>%
  ggplot(aes(x = "bp_systolic", y = bp_systolic)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```



```
data %>%
  ggplot(aes(x = "bp_diastolic", y = bp_diastolic)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```



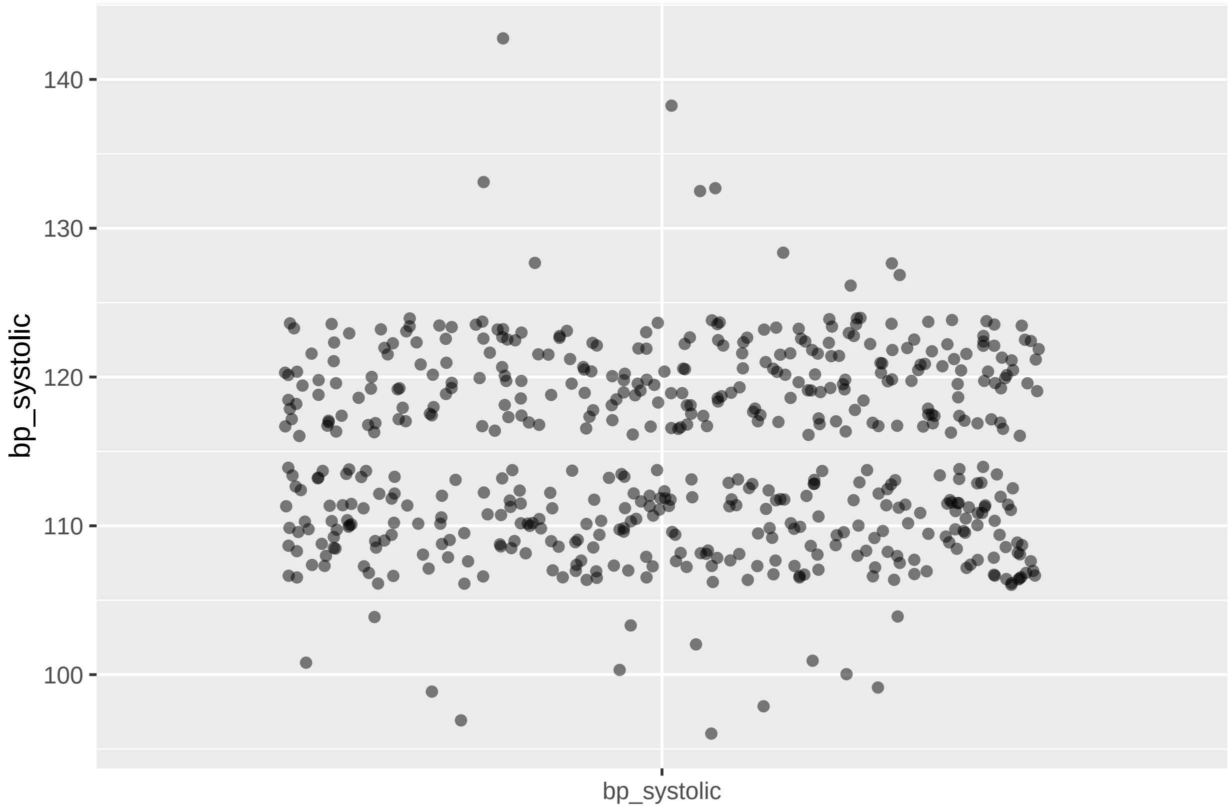
We can observe that there are two different atypical patients with a very odd blood presure. Both of them have a diastolic or systolic blood presure of almost 0 mm/Hg, which is impossible for a living human being. Then, both of them have been set to NA.

```
data %>%
  filter(bp_diastolic < 50 |bp_systolic < 50) %>%
  pull(id)
```

```
## [1] "162" "201"
```

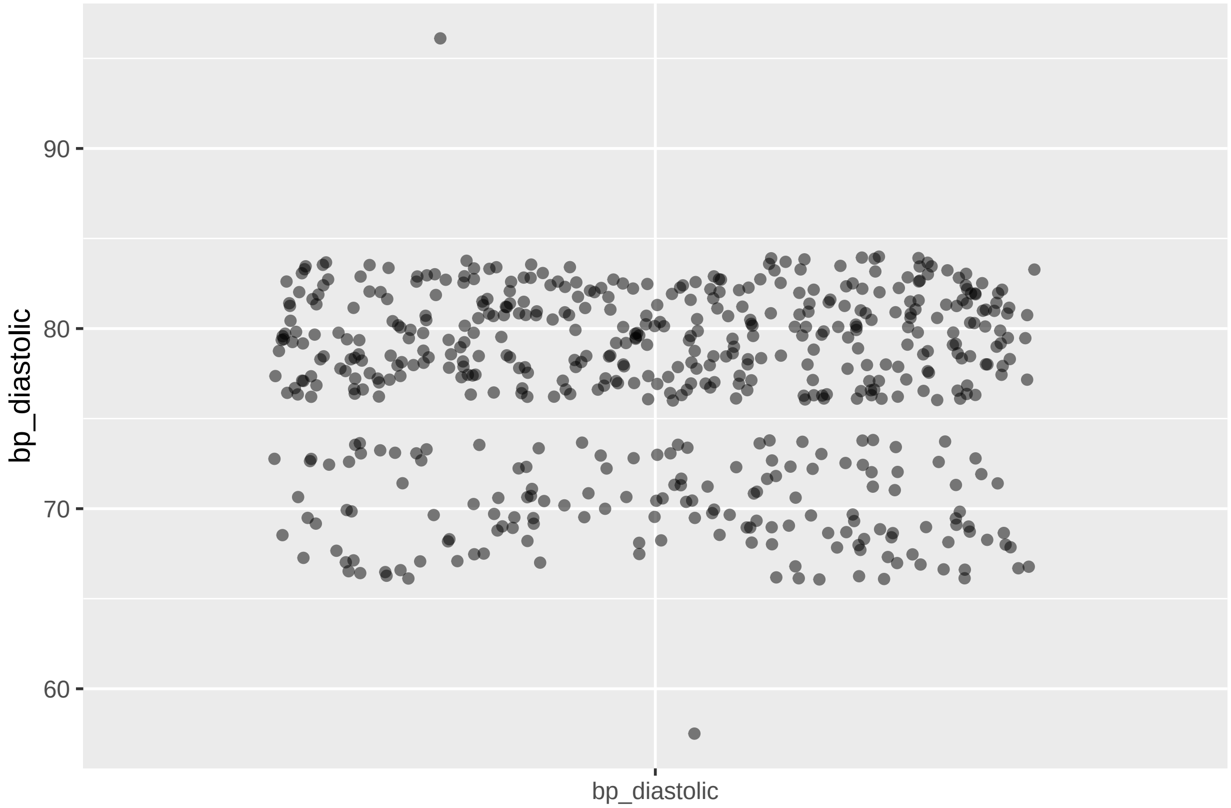
```
#I will flag this patient in case it pops out somewhere else in the analysis.
data[data$bp_diastolic < 15,"bp_diastolic" ] = NA
data[data$bp_systolic < 15,"bp_systolic" ] = NA

#Re plot the data
data %>%
  ggplot(aes(x = "bp_systolic", y = bp_systolic)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



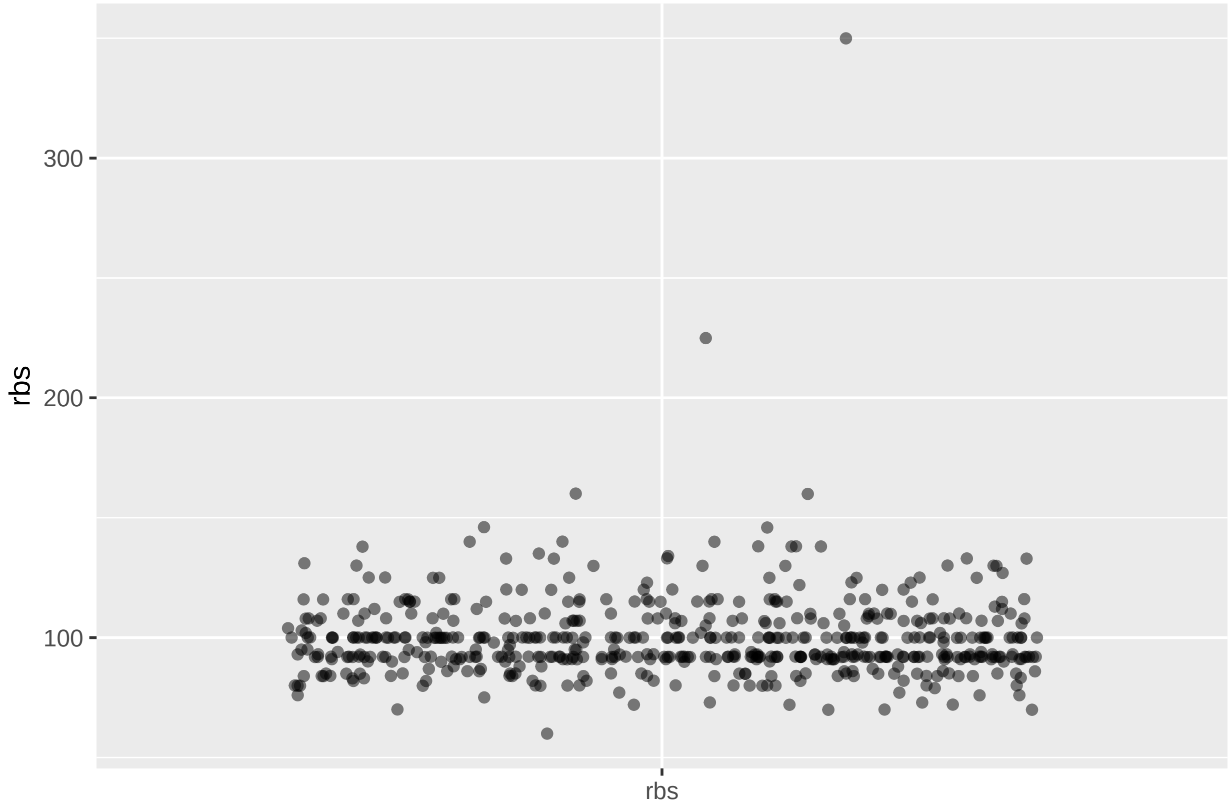
```r
data %>%
  ggplot(aes(x = "bp_diastolic", y = bp_diastolic)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



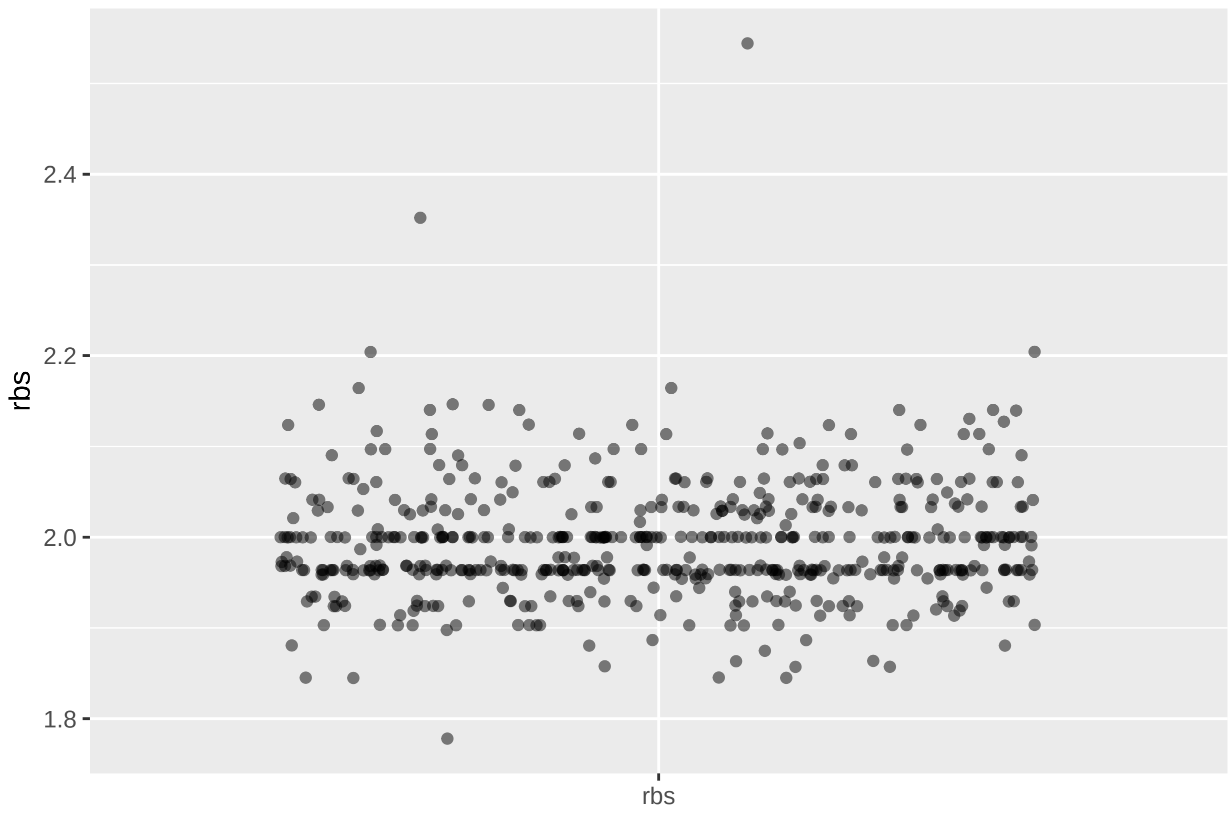
3.1.11. Random blood sugar (glucose) test Now, I will look for outliers in the random blood sugar (glucose) test.

```
data %>%
  ggplot(aes(x = "rbs", y = rbs)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```



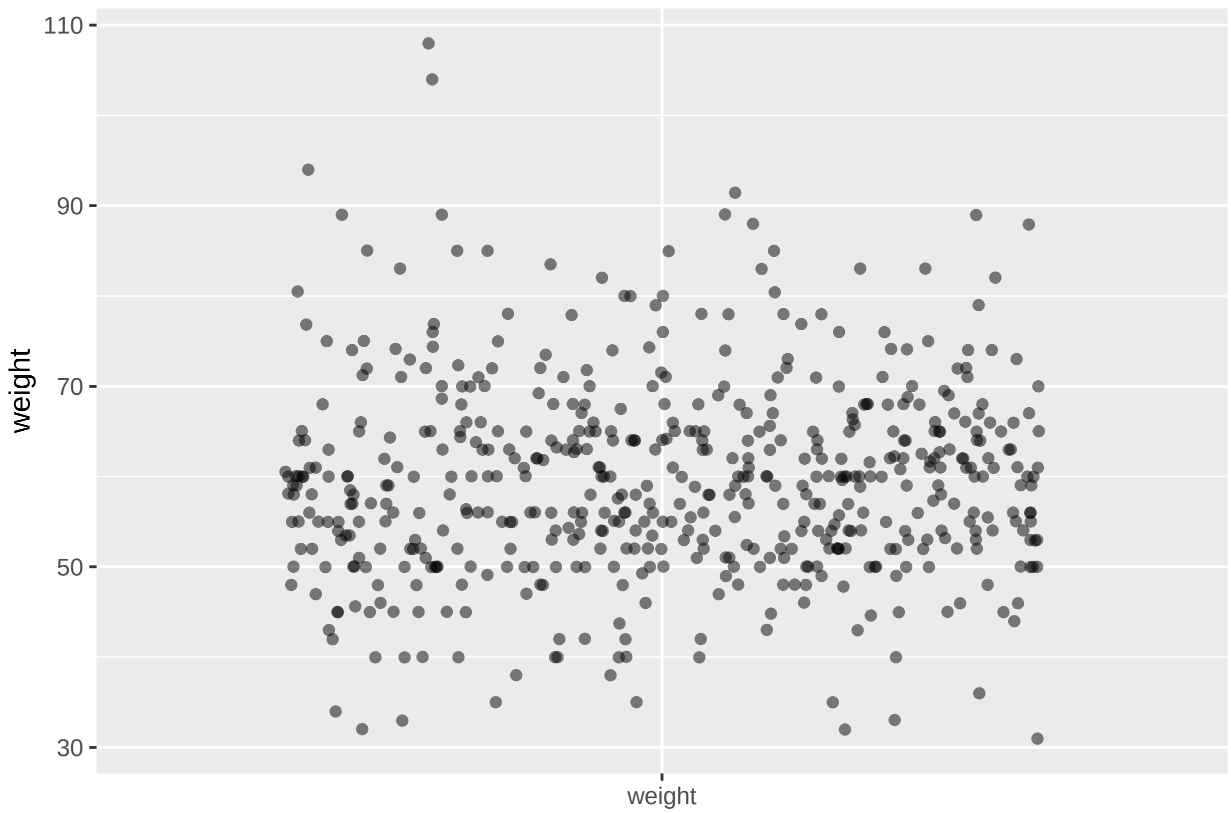
According to the literature, glucose levels can go as up as the ones that are observed. This would likely imply the existence of a syndrome, as well as many physiological consequences. Since this value is then biologically possible, it will be retained. However, the variable will be log-10 transformed.

```
#Transform the variable
data = data %>%
  mutate(rbs = log10(rbs))

#Re plot the data
data %>%
  ggplot(aes(x = "rbs", y = rbs)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```

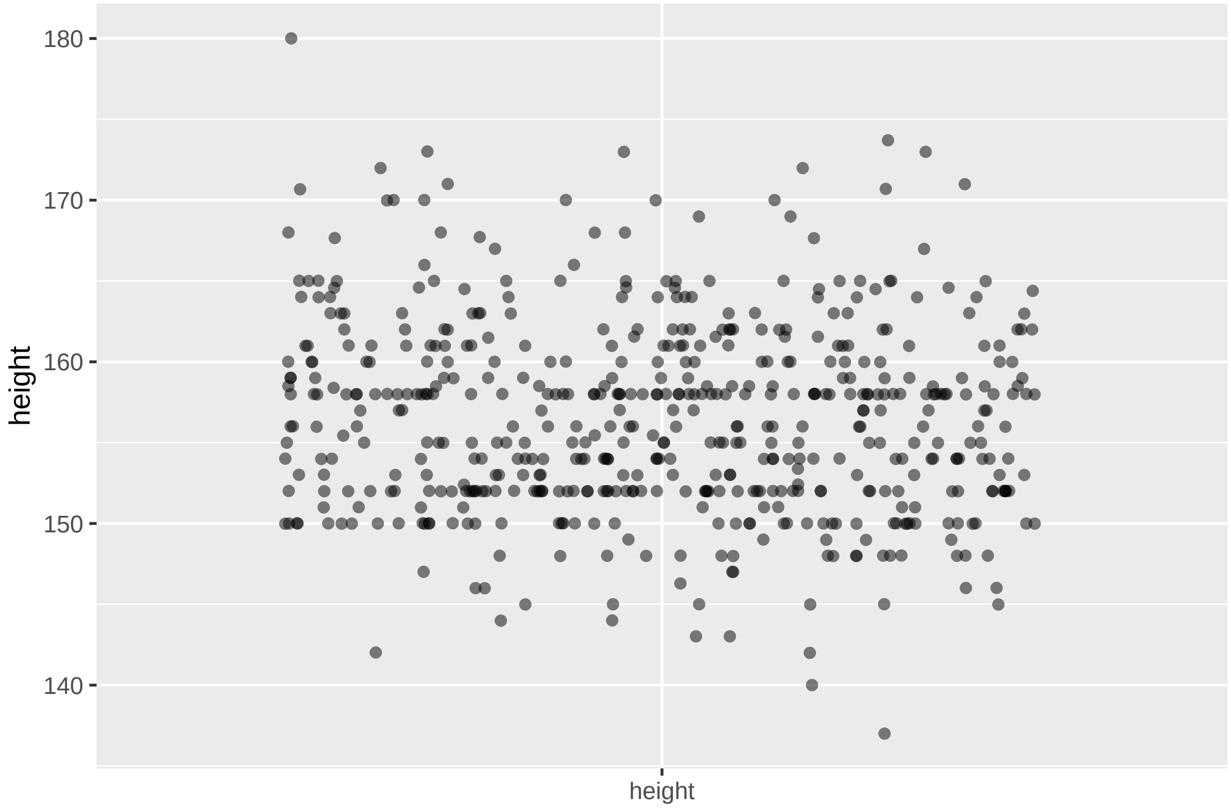


3.1.12. Other Other biological continuous variables were checked for outliers, but in view of no significant outliers, these variables were not further investigated or transformed.

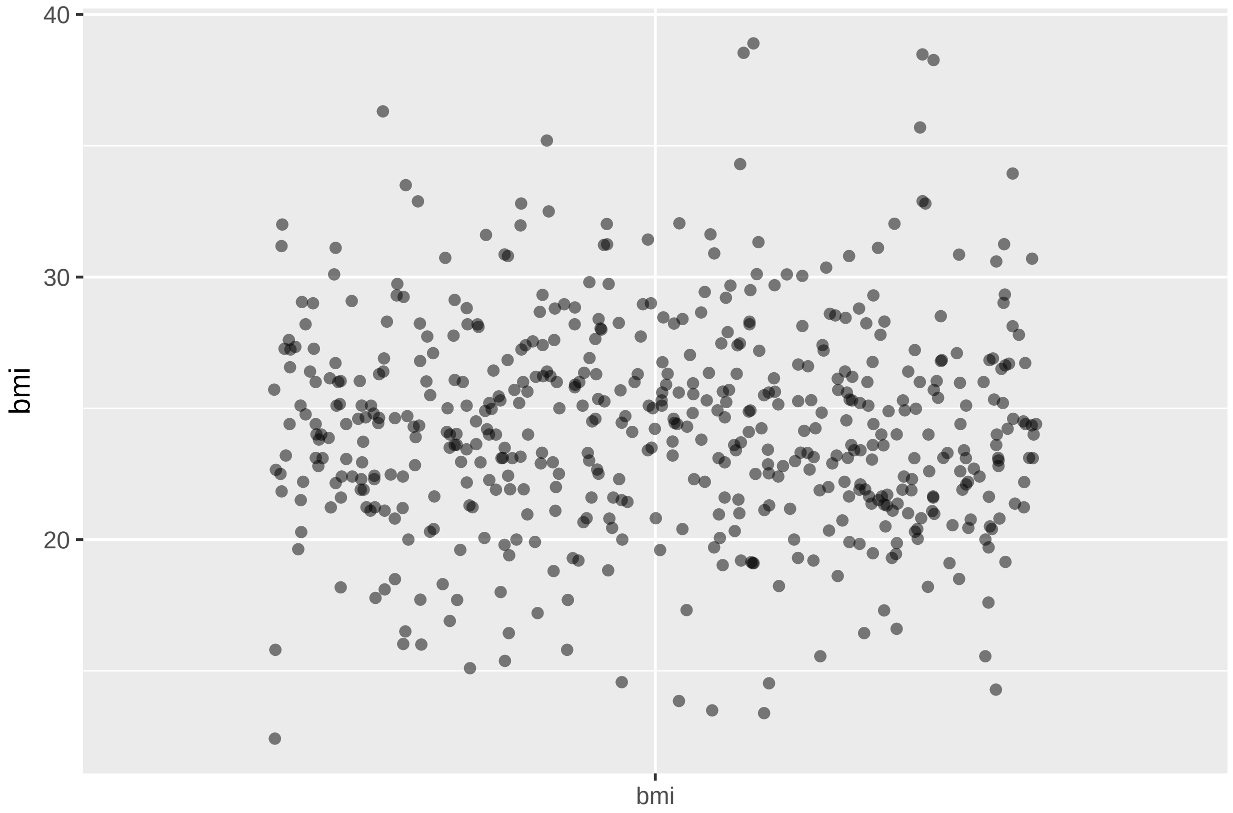
```r
# Plot to identify outliers for weight
data %>%
  ggplot(aes(x = "weight", y = weight)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```



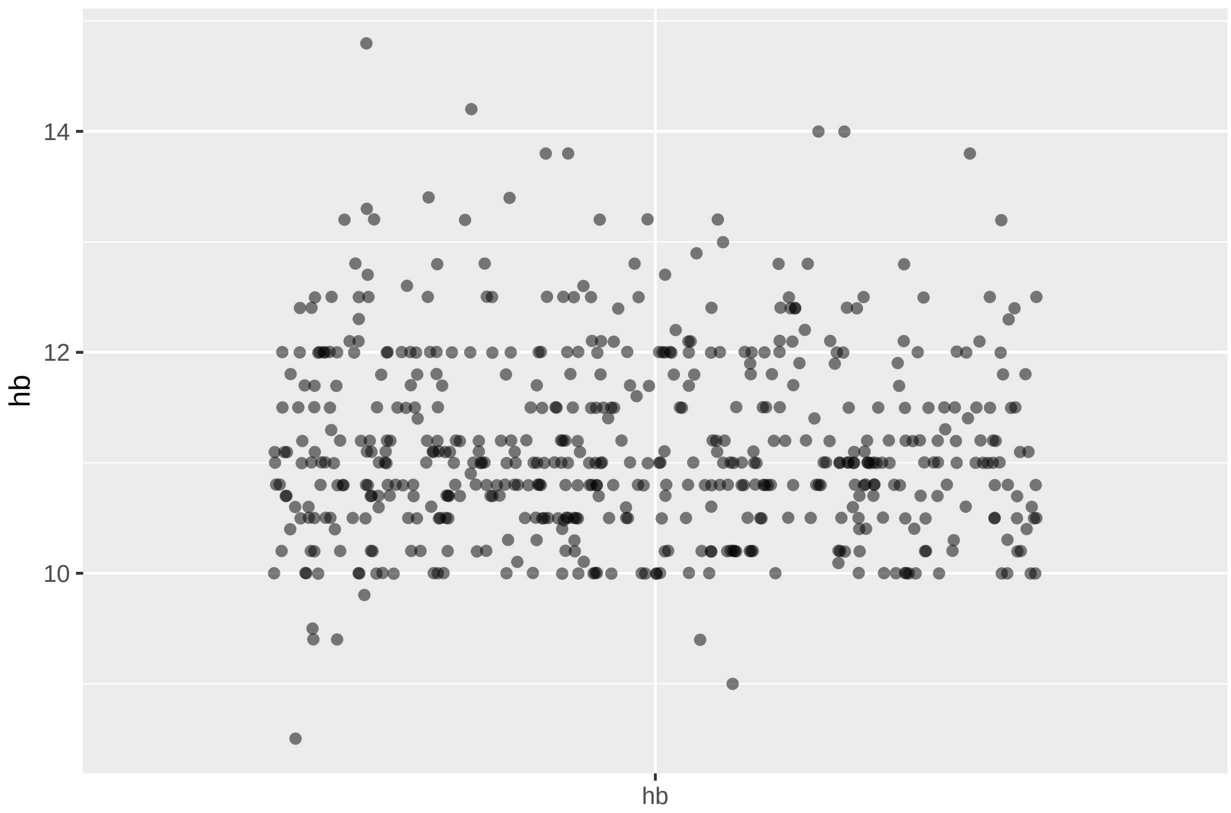
```r
# Plot to identify outliers for height
data %>%
  ggplot(aes(x = "height", y = height)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```



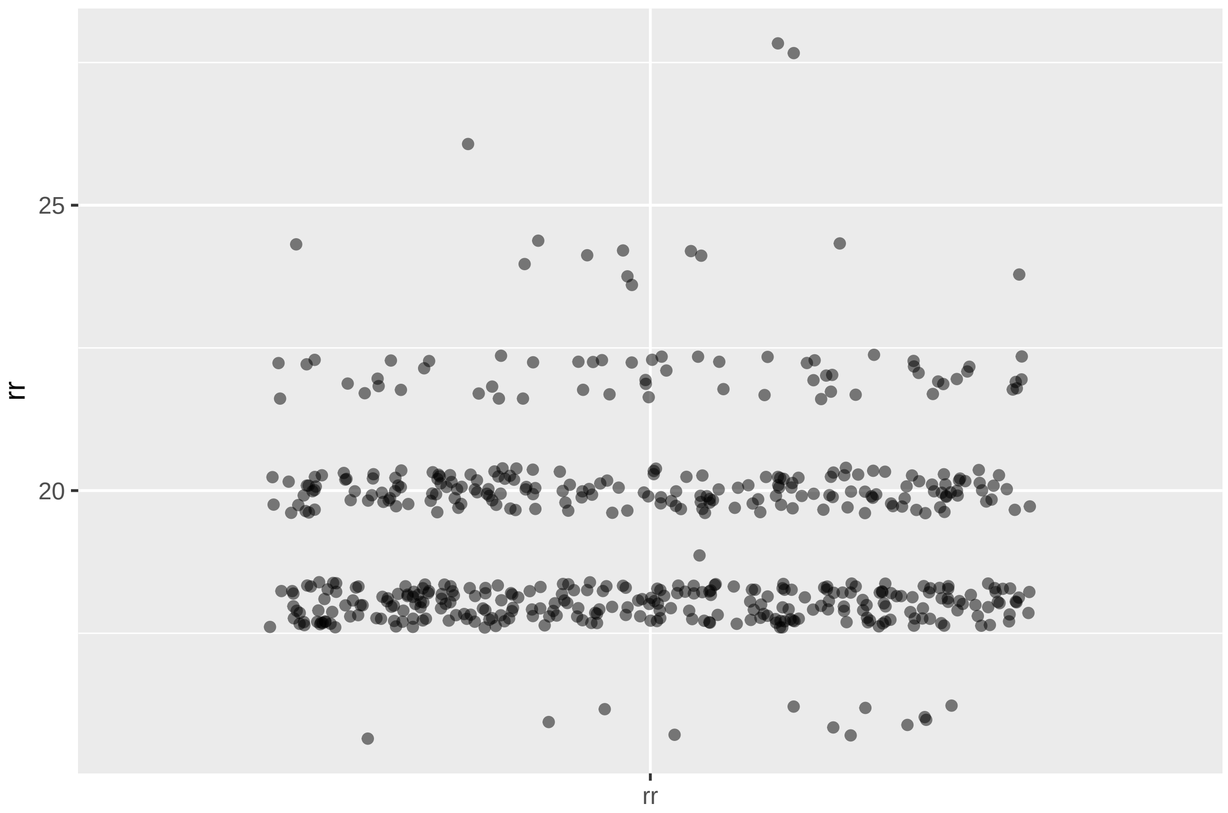
```
# Plot to identify outliers for blood mass index (BMI)
data %>%
  ggplot(aes(x = "bmi", y = bmi)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```



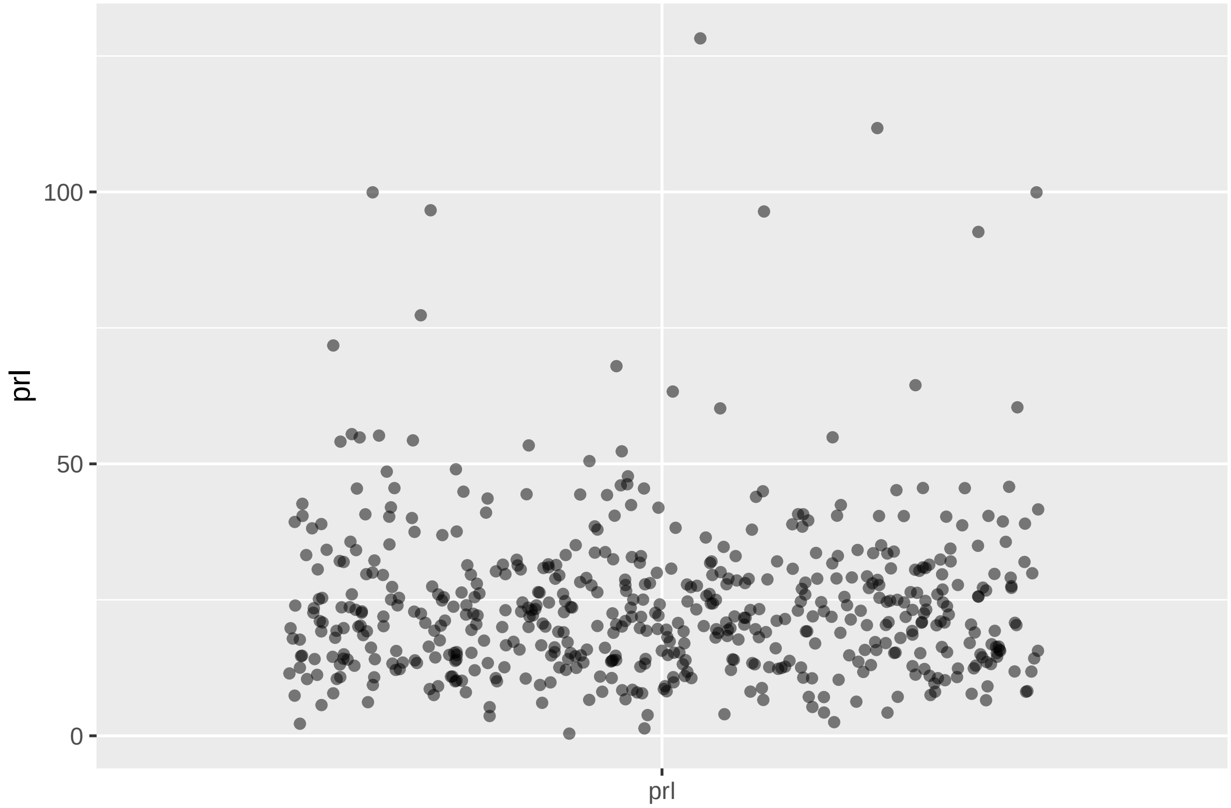
```
# Plot to identify outliers for hemoglobin (hb)
data %>%
  ggplot(aes(x = "hb", y = hb)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```



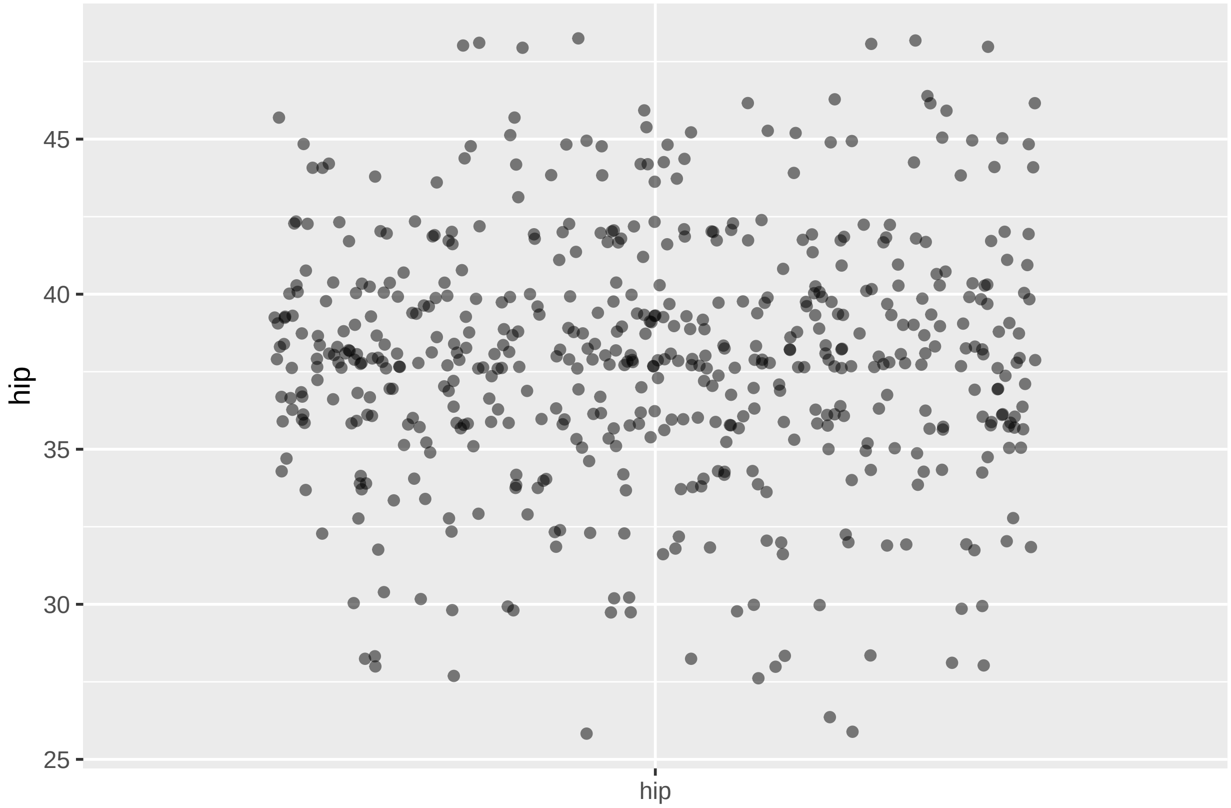
```
# Plot to identify outliers for respiratory rate (RR)
data %>%
  ggplot(aes(x = "rr", y = rr)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```



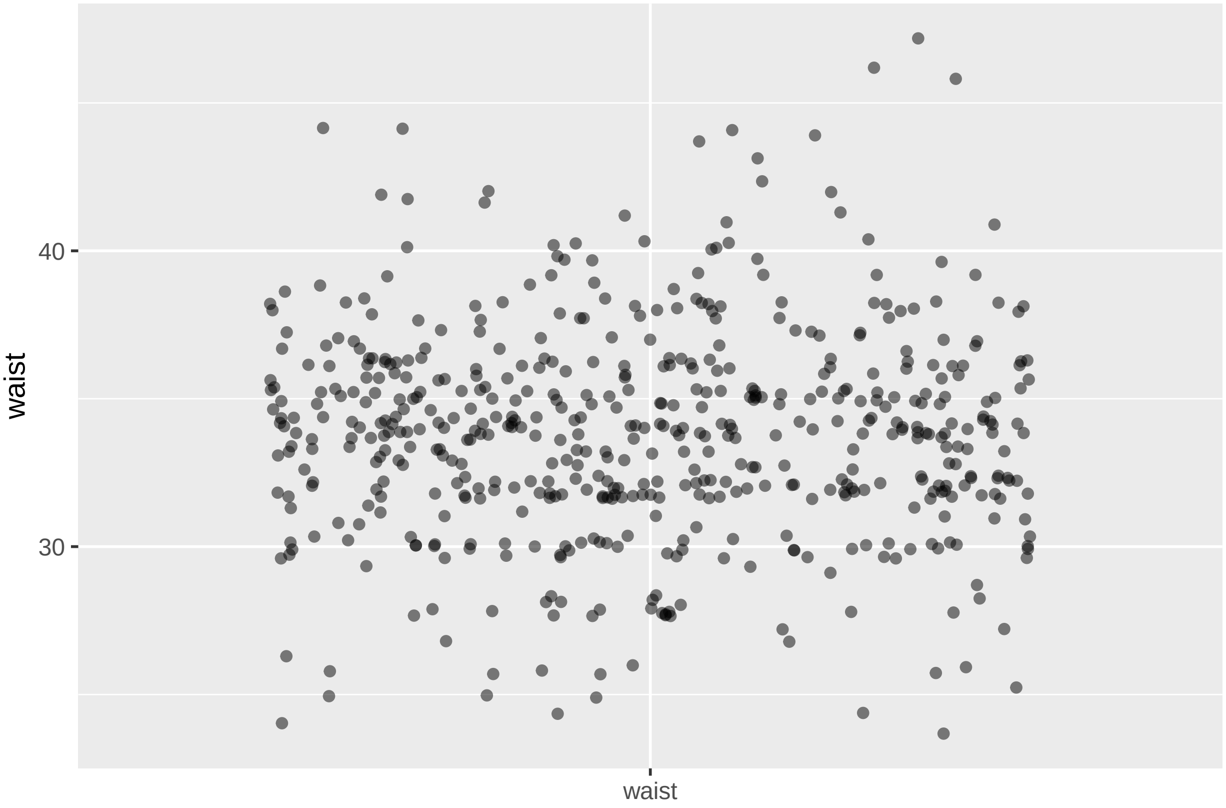
```
# Plot to identify outliers for prolactin (prl)
data %>%
  ggplot(aes(x = "prl", y = prl)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```



```
# Plot to identify outliers for hip circumference
data %>%
  ggplot(aes(x = "hip", y = hip)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```



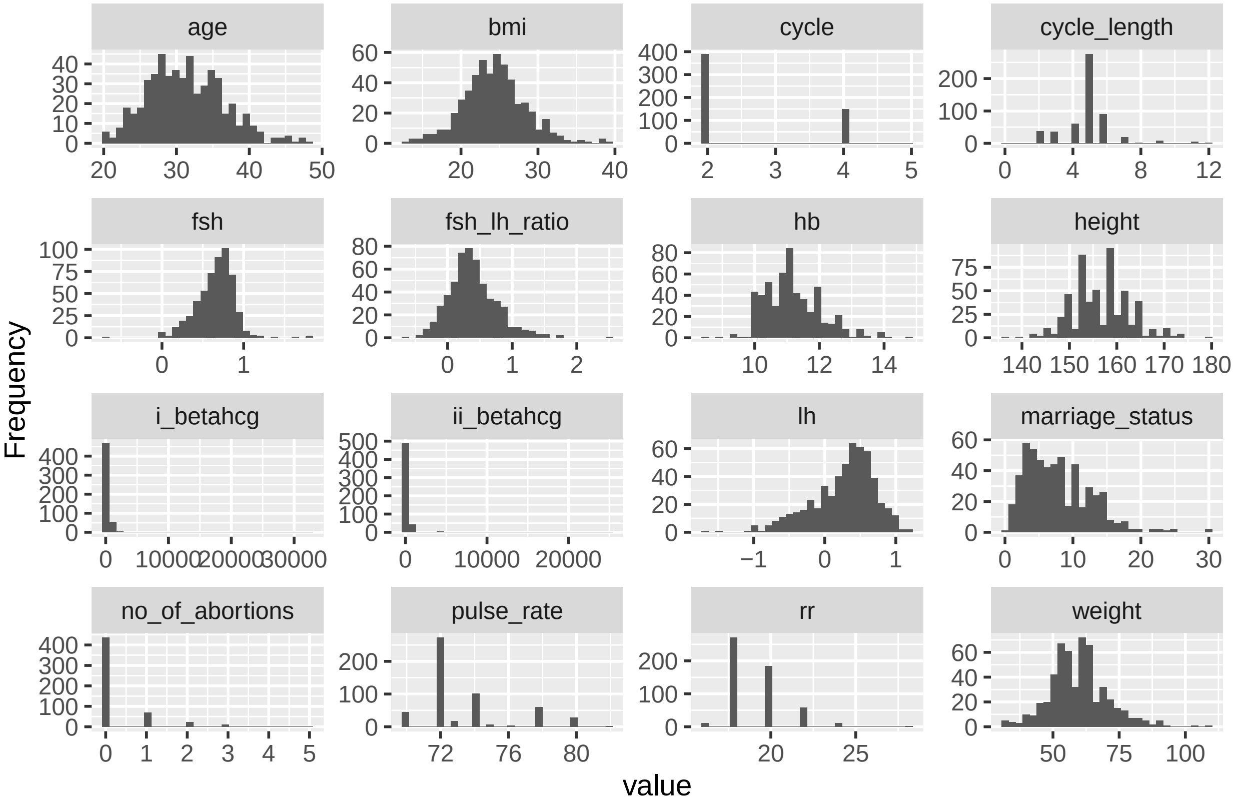
```
# Plot to identify outliers for waist circumference
data %>%
  ggplot(aes(x = "waist", y = waist)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```



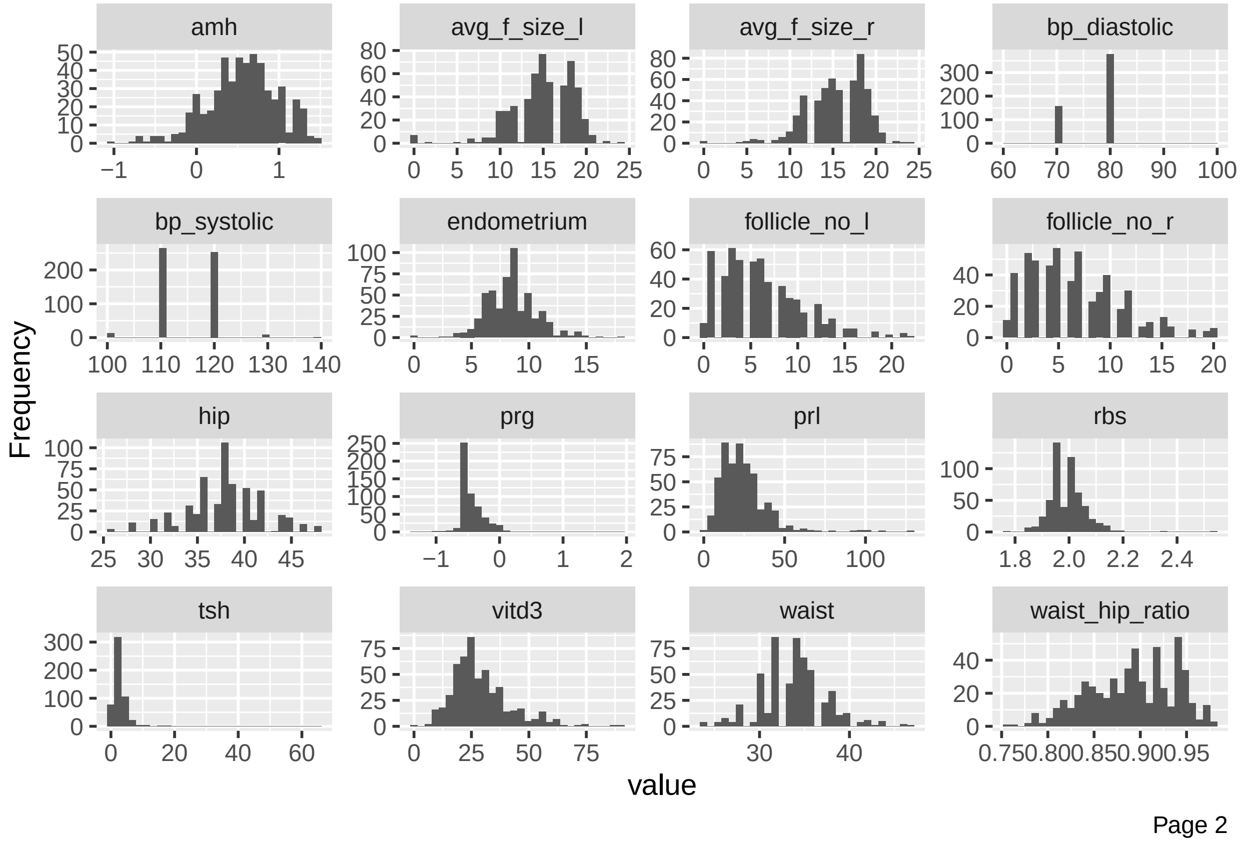
3.2 Variablity following data transformations and outlier analysis

The variables arer rerplotted below to see how the distribution of the variables has changed and the number of missing variables in the dataset, which now stands at 14 compared to the initial 3 missing values.

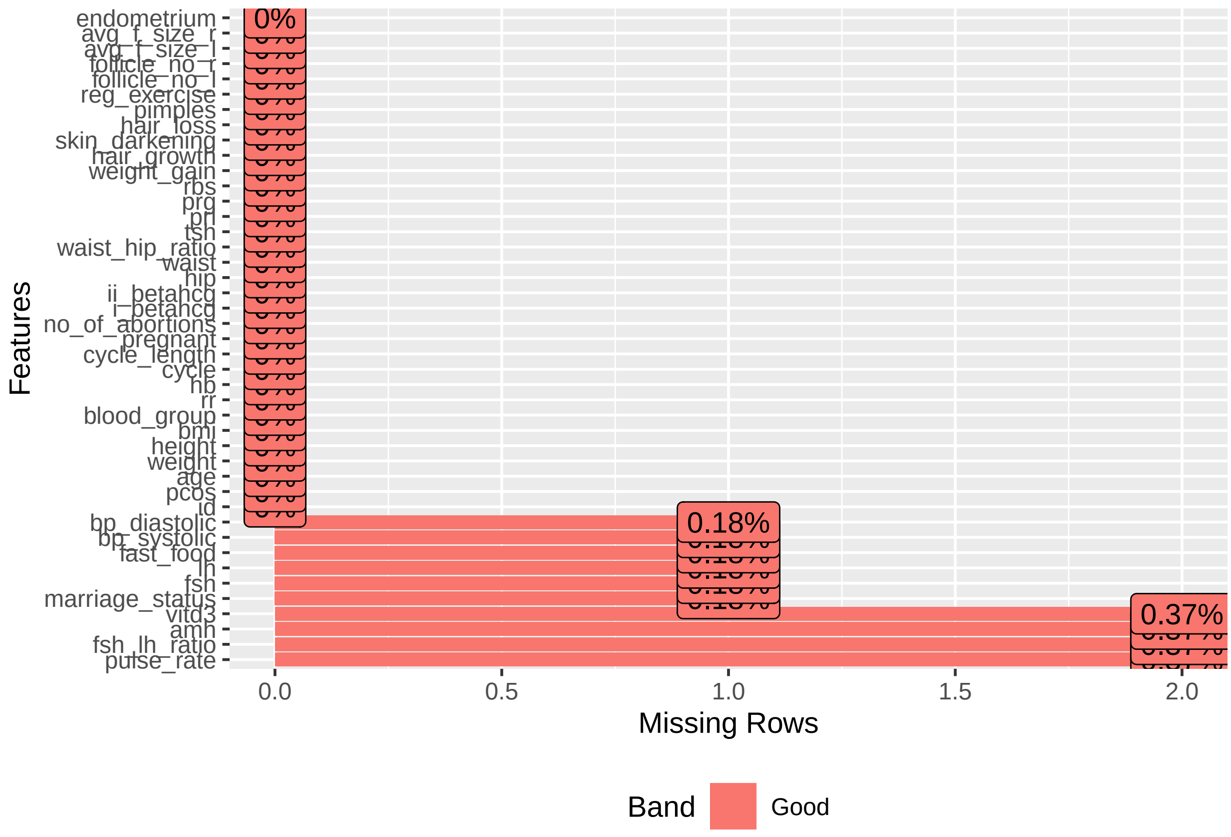
```
plot_histogram(data)
```


Frequency
value
age
0
10
20
30
40
20
30
40
50
bmi
0
20
40
60
20
30
40
cycle
0
100
200
300
400
2
3
4
5
cycle_length
0
100
200
0
4
8
12
fsh
0
25
50
75
100
0
1
fsh_lh_ratio
0
20
40
60
80
0
1
2
hb
0
20
40
60
80
10
12
14
height
0
25
50
75
140
150
160
170
180
i_betahcg
0
100
200
300
400
0
10000
20000
30000
ii_betahcg
0
100
200
300
400
500
0
10000
20000
lh
0
20
40
60
−1
0
1
marriage_status
0
20
40
60
0
10
20
30
no_of_abortions
0
100
200
300
400
0
1
2
3
4
5
pulse_rate
0
100
200
72
76
80
rr
0
100
200
20
25
weight
0
20
40
60
50
75
100

Page 1



```
plot_missing(data)
```



The code demonstrates that the missing data points are all for different individuals, thus supporting the assumption that they are missing randomly and may be random transcription errors.

```
data %>%
  filter(is.na(fast_food) |
           is.na(marriage_status) |
           is.na(amh) |
           is.na(fsh) |
           is.na(lh) |
           is.na(vitd3) |
           is.na(pulse_rate) |
           is.na(bp_diastolic) |
           is.na(bp_systolic)) %>%
  dplyr::select(c(id, fast_food, marriage_status, amh, fsh, lh, vitd3, pulse_rate, bp_diastolic, bp_sys
  knitr::kable()
```

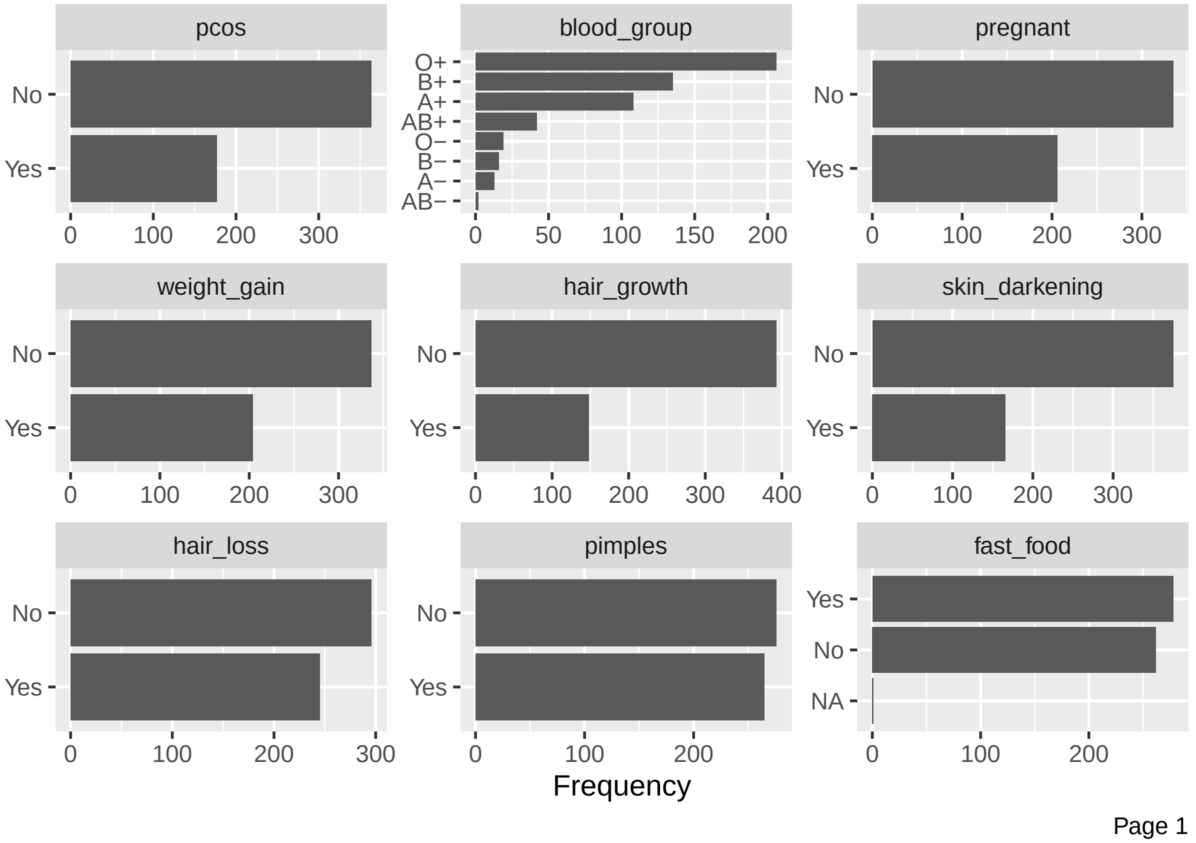
id	fast_food	marriage_status	amh	fsh	lh	vitd3	pulse_rate	bp_diastolic	bp_systolic
157	NA	5	0.7218106	0.4899585	-0.2291480	25.30	72	70	120
162	No	18	-0.3467875	0.3979400	0.4031205	24.00	75	80	NA
192	Yes	8	0.8068580	0.5599066	0.0086002	NA	74	70	120
196	Yes	14	0.8228216	1.3424227	0.5301997	NA	72	80	120
201	Yes	10	0.0170333	0.9537597	0.4913617	22.00	73	NA	120
224	Yes	5	0.8920946	0.8475727	0.6117233	31.80	NA	70	120

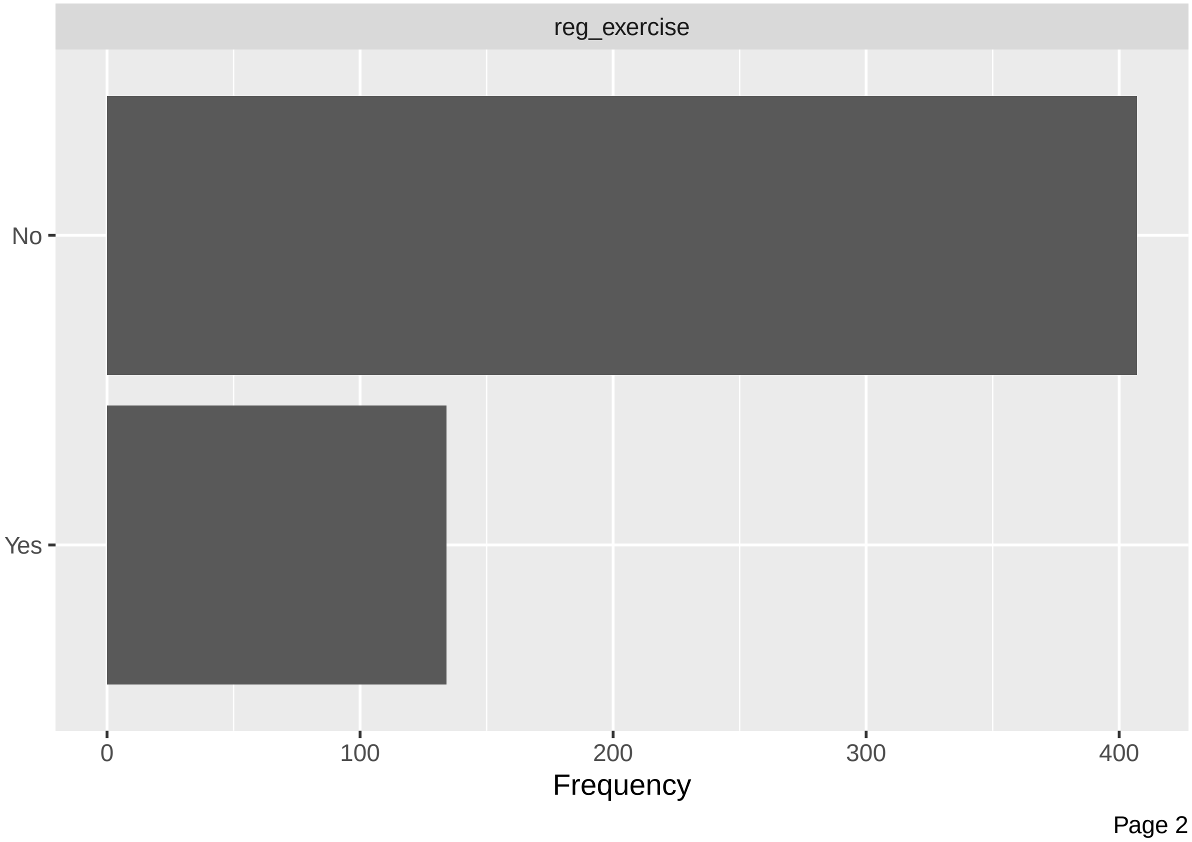
id	fast_food	marriage_status	amh	fsh	lh	vitd3	pulse_rate	bp_diastolic	bp_systolic
268	Yes	1	NA	0.8662873	0.5575072	30.20	72	80	120
297	No	12	1.0334238	0.8259451	0.7101174	24.90	NA	70	110
306	No	9	NA	0.4638930	-0.4559320	38.60	74	70	120
330	Yes	5	0.5440680	NA	0.5658478	28.60	72	80	110
456	No	12	0.8864907	0.6364879	NA	41.04	70	80	110
459	No	NA	0.8195439	0.2148438	-0.7695511	20.80	72	80	120

4. Variation in categorical variables

The variation of the categorical variables is displayed below.

```
plot_bar(data %>%
           dplyr::select(-id))
```



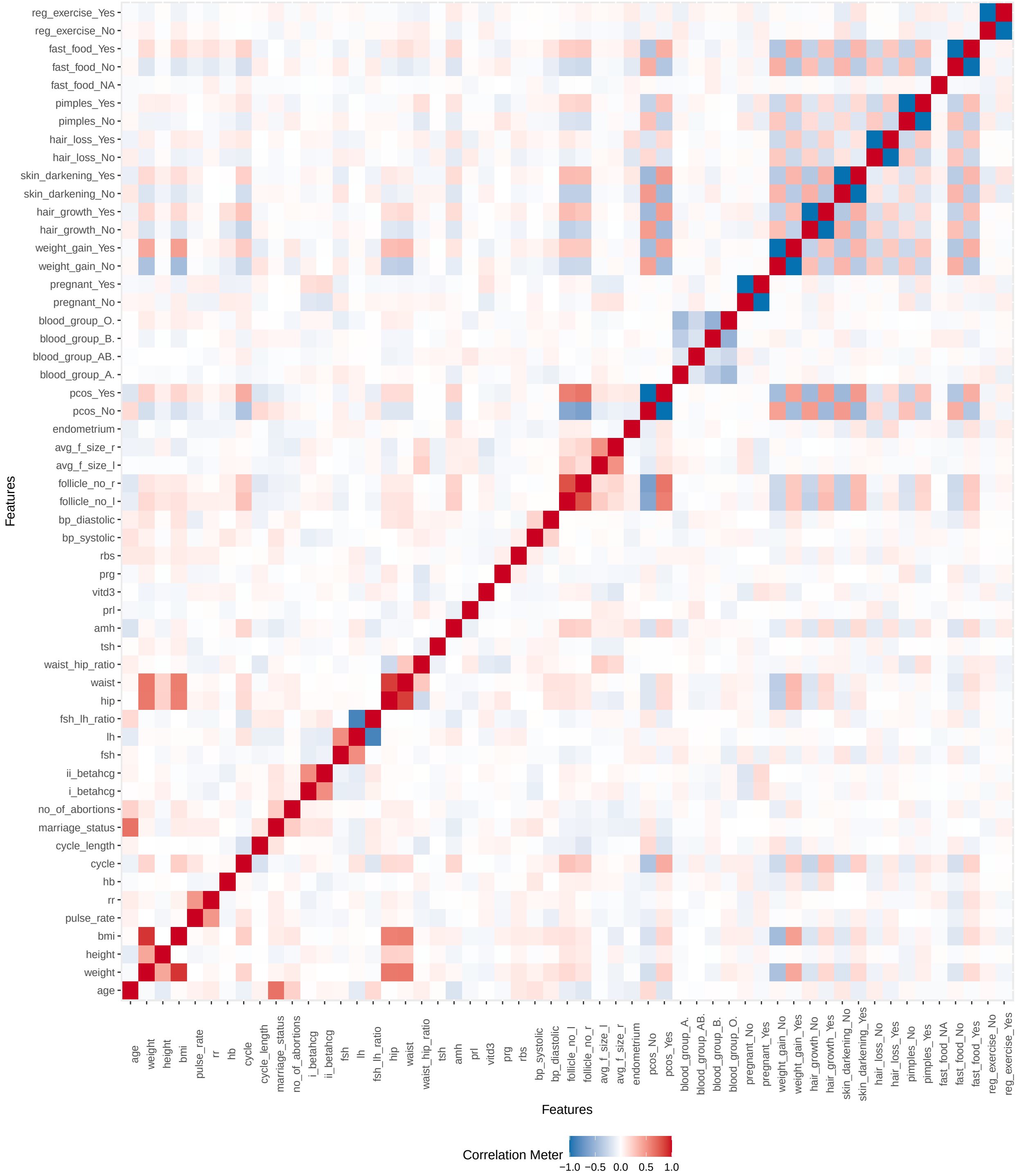


Importanly, we can observe a class imbalance in our response variable (diagnosis of PCOS) with 364 patients that arre negative for PCOS and 172 with a diagnosis of PCOS. Interestingly, presence of pimples, consumption of fast food and hair loss seem to be present and absent in almost an equal number of patients. None of the variables show an important lack of variation.

5. Correlations between variables

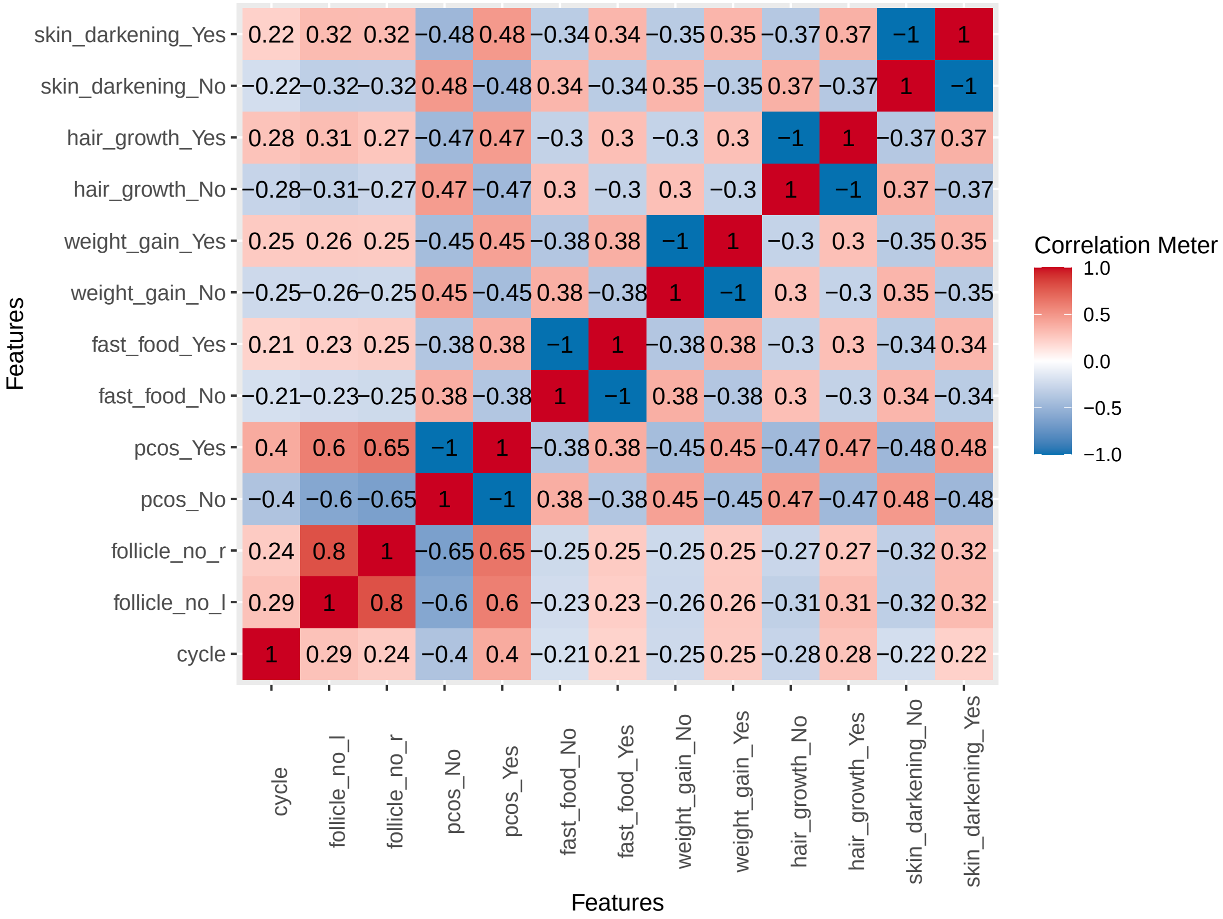
Finally, covariation of all of the variables in the dataset will be explore to see if there's any strong correlation that needs to be accounted for.

```
plot_correlation(data %>%
                   dplyr::select(-id),
                 type = 'all',
                 cor_args = list("use" = "complete.obs"))
```



The correlation plot reveals some variables that are directly correlated with our outcome of interest (i.e. diagnosis of PCOS). These include skin darkening, hair growth, weight gain, cycle length and number of follicles in each ovary. This is unsurprising, as these are all recognized features or diagnostic criteria for PCOS. However, fast food seems to also be correlated with PCOS. Whilst a change in diet to a more western diet of processed and fast food has been suggested to be involved in the increasing prevalence of PCOS, the physiological mechanism for this is not clear. Consumption of fast food is not often part of the diagnostic criteria or risk factors considered in the diagnosis of PCOS. The features least correlated with a diagnosis of PCOS seem to be blood group, pregnancy status, beta-HCG, and respiratory rate. This finding is expected as none of these variables have been linked to PCOS in the scientific literature. Some other interesting correlations can be gleamed from this plot. Namely that fast food is correlated with weight gain, hair growth, hair loss, pimples, skin darkening and follicle numbers in both ovaries, all of which are also correlated with a diagnosis of PCOS. This is an unexpected and interesting finding, and raises questions about the relationship between PCOS and the consumption of fast food. Finally, some expected correlations are noted. Pregnancy is strongly correlated with beta-HCG values, which are normally elevated in pregnancy. Hip and waist circumferences, which are measures used to denote central obesity, are correlated with weight and body mass index. Weight gain is similarly correlated with hip and waist circumference, weight and body mass index.

```
# Correlation plot for all variables that have a correlation of >0.35 with the outcome of interest

figure5.1 <- plot_correlation(data %>% na.omit(data) %>% dplyr::select(pcos, cycle, fast_food, weight_ga
                              type = 'all',
                              theme_config = theme(legend.position = "right", axis.text = element_text(
                              cor_args = list("use" = "complete.obs"))
```



6. Associations between variables

6.1 Variables excluded from the data used for models

Several variables have not been included in the data used to build our predictive diagnostic models for PCOS. These were parameters related to infertility (pregnancy, beta human chorionic gonadotropin levels) and those without clear evidence of association with PCOS (marital status, blood group, thyroid stimulating hormone levels, respiratory and pulse rate, hemoglobin). In this section, we explore the association of these variables with a diagnosis of PCOS (i.e. outcome of interest) to further justify their removal from the machine learning models.

Two additional packages *epiDisplay* and *gmodels* need to be loaded (and installed if not done previously) to visualize the associations.

```
library(epiDisplay)
library(gmodels)
```

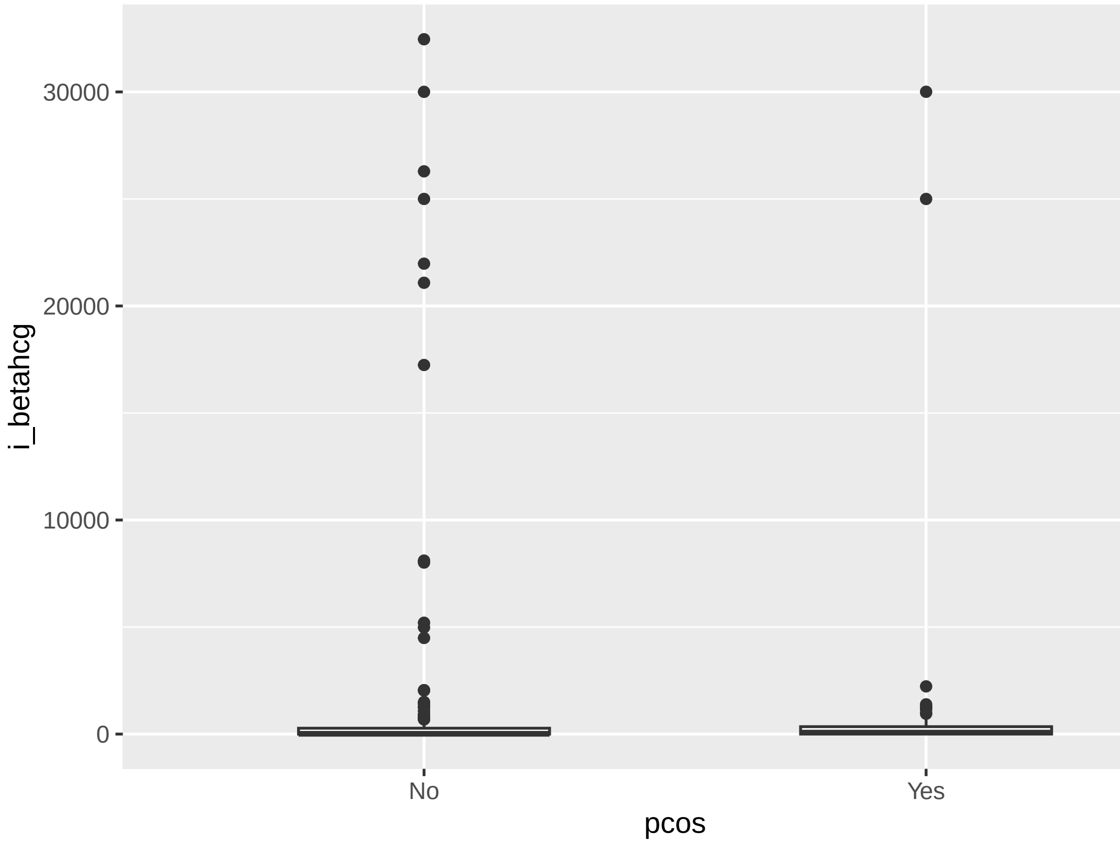
```

# Table and chi squared test to summarize association between pregnancy and diagnosis of PCOS
table6.1.1 <- CrossTable(data$pregnant, data$pcos, prop.t=FALSE, prop.r=FALSE, prop.c=TRUE, expected = 
```

6.1.1 Pregnancy

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  541
##
##
##              | PCOS Diagnosis
##     Pregnant |        No |       Yes | Row Total |
## -------------|-----------|-----------|-----------|
##           No |       222 |       113 |       335 |
##              |     0.610 |     0.638 |           |
## -------------|-----------|-----------|-----------|
##          Yes |       142 |        64 |       206 |
##              |     0.390 |     0.362 |           |
## -------------|-----------|-----------|-----------|
## Column Total |       364 |       177 |       541 |
##              |     0.673 |     0.327 |           |
## -------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  0.4110566     d.f. =  1     p =  0.5214337
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 =  0.2989687     d.f. =  1     p =  0.5845297
##
##
```

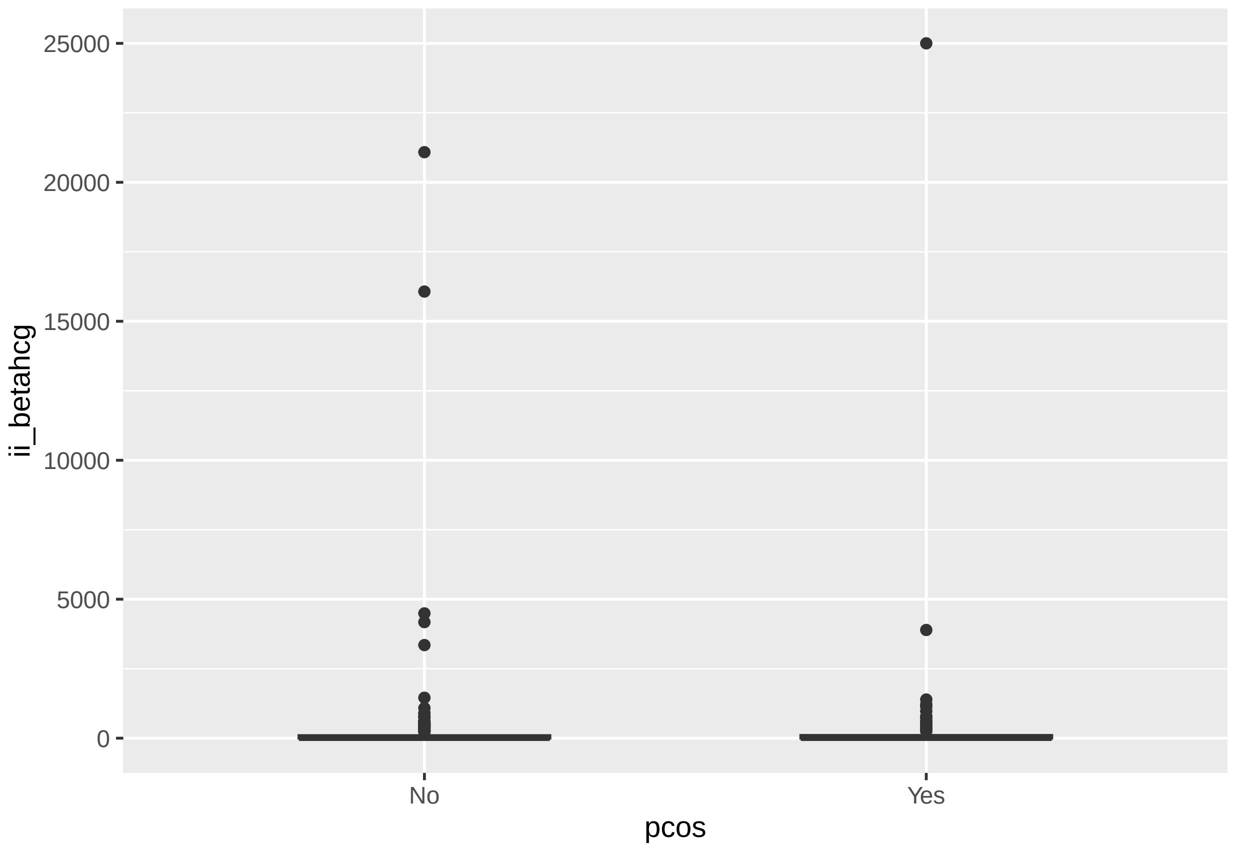
There is no statistically significant difference between patients with and without PCOS with regards to pregnancy status (p = 0.521). This further supports the decision to not use this variable in the predictive models.

```
# Visualize beta-HCG levels between patients with and without PCOS for both types of beta-HCG test
ggplot(data, aes(pcos, i_betahcg)) + geom_boxplot(width = 0.5)
```



6.1.2 beta-HCG levels

```
ggplot(data, aes(pcos, ii_betahcg)) + geom_boxplot(width = 0.5)
```



```
# Mean and standard deviation by PCOS diagnosis for both types of beta-HCG test
data %>% dplyr::select(i_betahcg, pcos) %>% group_by(pcos) %>%
  summarise(n = n(),
            mean = mean(i_betahcg, na.rm = TRUE),
            sd = sd(i_betahcg, na.rm = TRUE))
```

```
## # A tibble: 2 x 4
##   pcos      n  mean    sd
##   <fct> <int> <dbl> <dbl>
## 1 No      364  729. 3540.
## 2 Yes     177  532. 2923.
```

```
data %>% dplyr::select(ii_betahcg, pcos) %>% group_by(pcos) %>%
  summarise(n = n(),
            mean = mean(ii_betahcg, na.rm = TRUE),
            sd = sd(ii_betahcg, na.rm = TRUE))
```

```
## # A tibble: 2 x 4
##   pcos      n  mean    sd
##   <fct> <int> <dbl> <dbl>
## 1 No      364  224. 1437.
## 2 Yes     177  268. 1906.
```

```r
# Mean and standard deviation overall for both types of beta-HCG test
data %>% dplyr::select(i_betahcg) %>%
  summarise(n = n(),
            mean = mean(i_betahcg, na.rm = TRUE),
            sd = sd(i_betahcg, na.rm = TRUE))
```

```
## # A tibble: 1 x 3
##       n  mean    sd
##   <int> <dbl> <dbl>
## 1   541  665. 3349.
```

```r
data %>% dplyr::select(ii_betahcg) %>%
  summarise(n = n(),
            mean = mean(ii_betahcg, na.rm = TRUE),
            sd = sd(ii_betahcg, na.rm = TRUE))
```

```
## # A tibble: 1 x 3
##       n  mean    sd
##   <int> <dbl> <dbl>
## 1   541  238. 1604.
```

```r
# Independent t-test to determine if there is a difference in beta-HCG levels between PCOS negative and
t.test(i_betahcg ~ pcos, data = data)
```

```
##
##  Welch Two Sample t-test
##
## data:  i_betahcg by pcos
## t = 0.68488, df = 414.36, p-value = 0.4938
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -368.3045  762.1843
## sample estimates:
##  mean in group No mean in group Yes
##          728.9824          532.0425
```

```r
t.test(ii_betahcg ~ pcos, data = data)
```

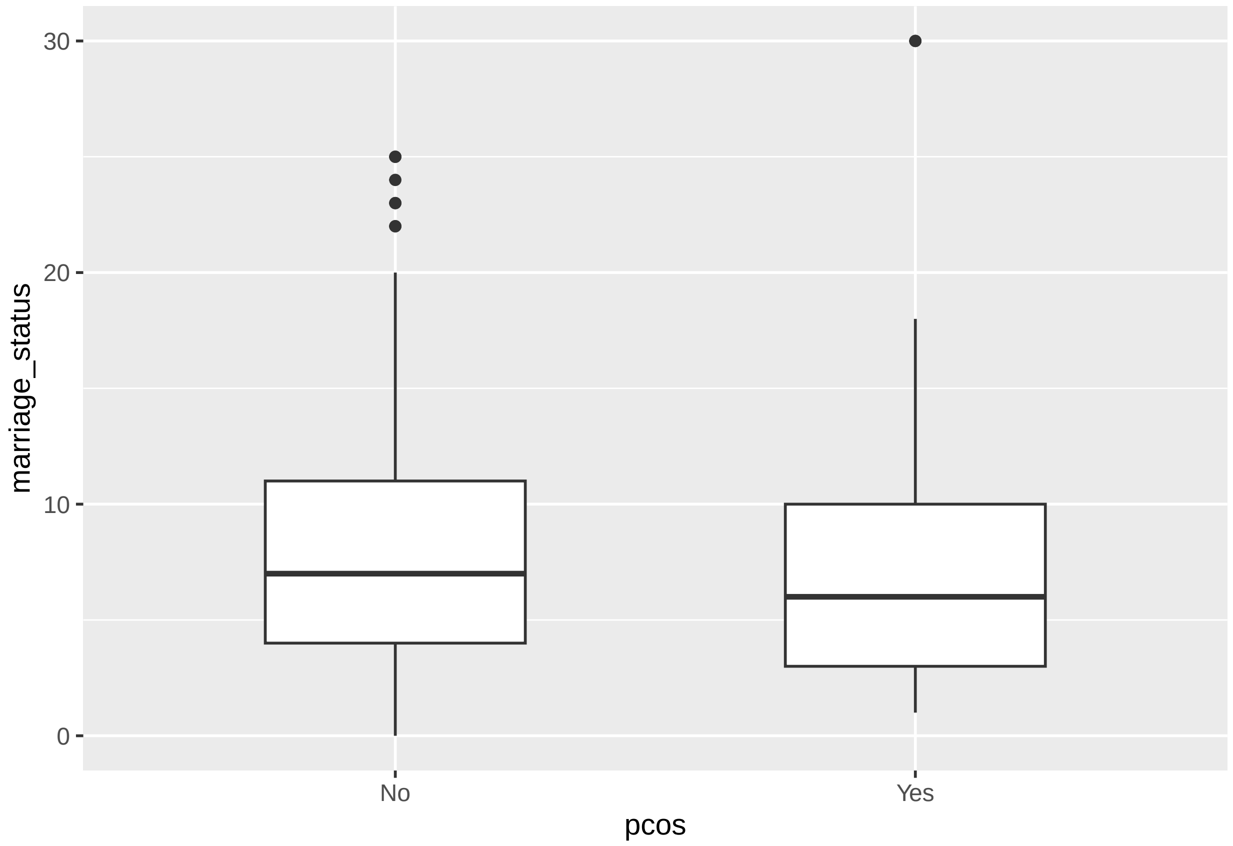
```
##
##  Welch Two Sample t-test
##
## data:  ii_betahcg by pcos
## t = -0.26928, df = 276.53, p-value = 0.7879
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -362.1692  275.0111
## sample estimates:
##  mean in group No mean in group Yes
##          223.9751          267.5542
```

There is no statistically significant difference between patients with and without PCOS with regards to beta-HCG levels (using either test i or ii) ($p = 0.4938$ for test i, $p = 0.7879$ for test ii). This further supports the decision to not use these variables in the predictive models.

```
# Visualize the length of marriage in years between patients with and without PCOS
ggplot(data, aes(pcos, marriage_status)) + geom_boxplot(width = 0.5)
```

6.1.3 Length of marriage

```
## Warning: Removed 1 rows containing non-finite values (`stat_boxplot()`).
```



```
# Mean and standard deviation by PCOS diagnosis for length of marriage in years
data %>% dplyr::select(marriage_status, pcos) %>% group_by(pcos) %>%
  summarise(n = n(),
            mean = mean(marriage_status, na.rm = TRUE),
            sd = sd(marriage_status, na.rm = TRUE))
```

```
## # A tibble: 2 x 4
##   pcos      n  mean    sd
##   <fct> <int> <dbl> <dbl>
## 1 No      364  8.06  4.82
## 2 Yes     177  6.90  4.70
```

```
# Mean and standard deviation overall for length of marriage in years
data %>% dplyr::select(marriage_status) %>%
  summarise(n = n(),
            mean = mean(marriage_status, na.rm = TRUE),
            sd = sd(marriage_status, na.rm = TRUE))
```

```
## # A tibble: 1 x 3
##       n  mean    sd
##   <int> <dbl> <dbl>
## 1   541  7.68  4.80
```

```
# Independent t-test to determine if there is a difference in length of marriage in years between PCOS
t.test(marriage_status ~ pcos, data = data)
```

```
##
##  Welch Two Sample t-test
##
## data:  marriage_status by pcos
## t = 2.6586, df = 354.04, p-value = 0.008203
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  0.3008461 2.0111294
## sample estimates:
##  mean in group No mean in group Yes
##          8.057692          6.901705
```

There is a statistically significant difference between patients with and without PCOS with regards to length of marriage (p = 0.008203). However, as this association does not have any basis in the pathophysiology of PCOS, this variable will still not be used in the predictive models.

```
# Table and chi squared test to summarize association between blood group and diagnosis of PCOS
table6.1.4 <- CrossTable(data$blood_group, data$pcos, prop.t=FALSE, prop.r=FALSE, prop.c=TRUE, expected
```

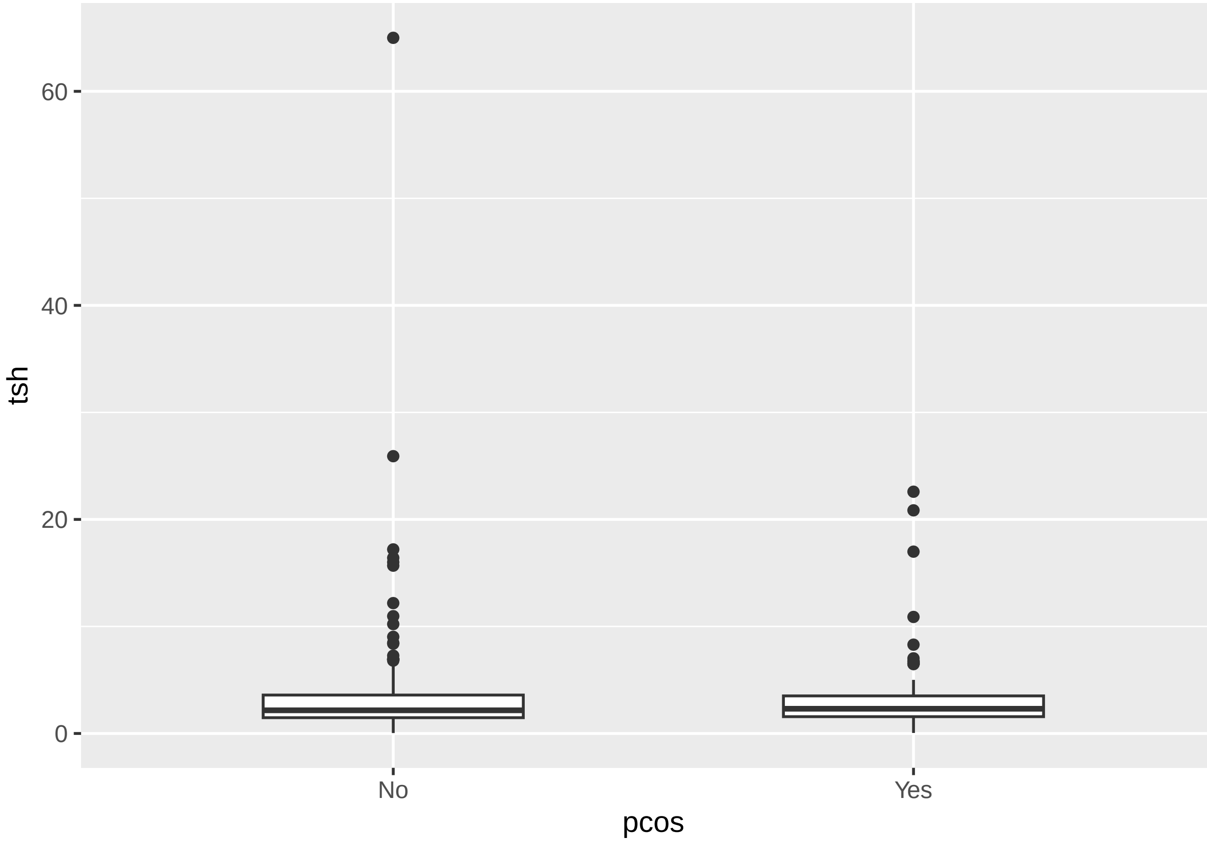
6.1.4 Blood group

```
## Warning in chisq.test(t, correct = FALSE, ...): Chi-squared approximation may be
## incorrect
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  541
##
##
##              | PCOS Diagnosis
##  Blood Group |        No |       Yes | Row Total | 
## -------------|-----------|-----------|-----------|
##           A+ |        74 |        34 |       108 | 
##              |     0.203 |     0.192 |           | 
## -------------|-----------|-----------|-----------|
##           A- |         9 |         4 |        13 | 
##              |     0.025 |     0.023 |           | 
## -------------|-----------|-----------|-----------|
##           B+ |        93 |        42 |       135 | 
##              |     0.255 |     0.237 |           | 
## -------------|-----------|-----------|-----------|
##           B- |        10 |         6 |        16 | 
##              |     0.027 |     0.034 |           | 
## -------------|-----------|-----------|-----------|
##           O+ |       140 |        66 |       206 | 
##              |     0.385 |     0.373 |           | 
## -------------|-----------|-----------|-----------|
##           O- |        11 |         8 |        19 | 
##              |     0.030 |     0.045 |           | 
## -------------|-----------|-----------|-----------|
##          AB+ |        26 |        16 |        42 | 
##              |     0.071 |     0.090 |           | 
## -------------|-----------|-----------|-----------|
##          AB- |         1 |         1 |         2 | 
##              |     0.003 |     0.006 |           | 
## -------------|-----------|-----------|-----------|
## Column Total |       364 |       177 |       541 | 
##              |     0.673 |     0.327 |           | 
## -------------|-----------|-----------|-----------|
## 
## 
## Statistics for All Table Factors
## 
## 
## Pearson's Chi-squared test 
## ------------------------------------------------------------
## Chi^2 =  2.048798     d.f. =  7     p =  0.9570902 
## 
## 
## 
## Fisher's Exact Test for Count Data
## ------------------------------------------------------------
## Alternative hypothesis: two.sided
## p =  0.9321499 
## 
## 
```

There is no statistically significant difference between patients with and without PCOS with regards to pregnancy status (p = 0.932). This further supports the decision to not use this variable in the predictive models.

```
# Visualize the TSH level between patients with and without PCOS
ggplot(data, aes(pcos, tsh)) + geom_boxplot(width = 0.5)
```



6.1.5 TSH level

```
# Mean and standard deviation by PCOS diagnosis for TSH level
data %>% dplyr::select(tsh, pcos) %>% group_by(pcos) %>%
  summarise(n = n(),
            mean = mean(tsh, na.rm = TRUE),
            sd = sd(tsh, na.rm = TRUE))
```

```
## # A tibble: 2 x 4
##   pcos      n  mean    sd
##   <fct> <int> <dbl> <dbl>
## 1 No      364  3.01  4.14
## 2 Yes     177  2.93  2.82
```

```
# Mean and standard deviation overall for TSH level
data %>% dplyr::select(tsh) %>%
  summarise(n = n(),
            mean = mean(tsh, na.rm = TRUE),
            sd = sd(tsh, na.rm = TRUE))
```

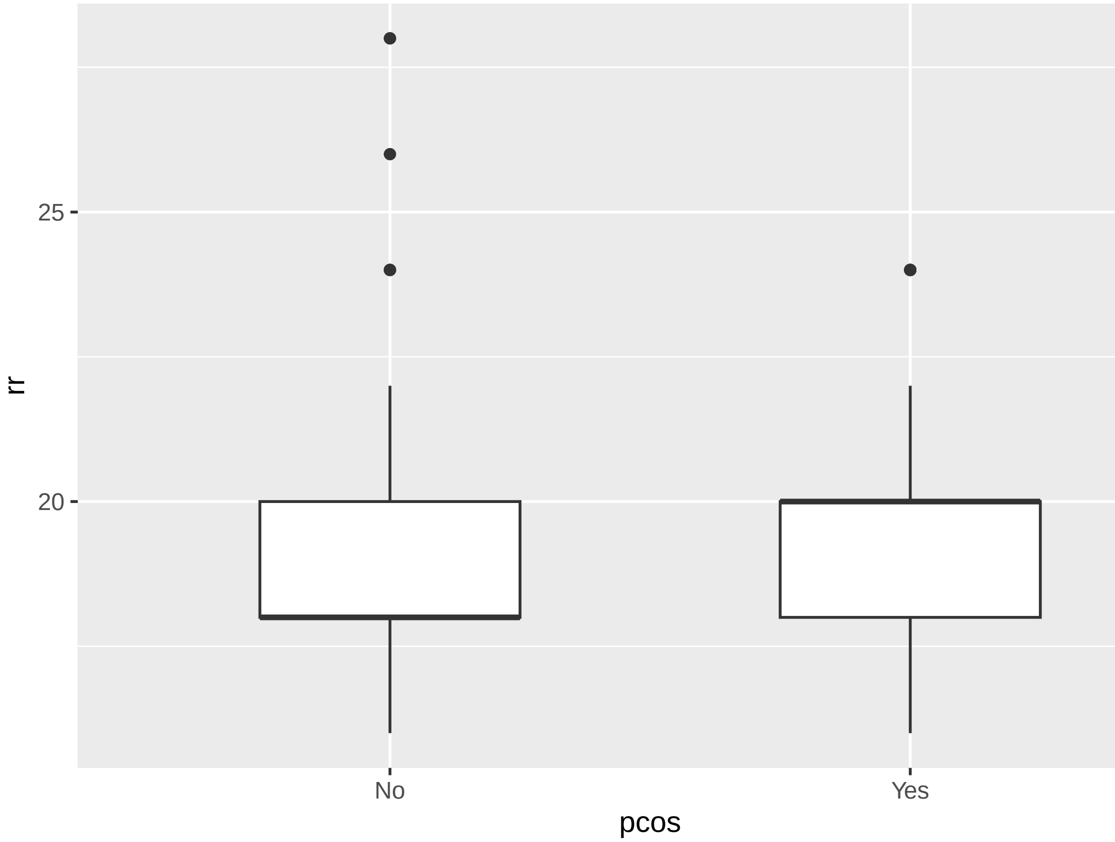
```
## # A tibble: 1 x 3
##       n  mean    sd
##   <int> <dbl> <dbl>
## 1   541  2.98  3.76
```

```
# Independent t-test to determine if there is a difference in TSH levels between PCOS negative and posi
t.test(tsh ~ pcos, data = data)
```

```
##
##  Welch Two Sample t-test
##
## data:  tsh by pcos
## t = 0.26727, df = 481.1, p-value = 0.7894
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -0.5150585  0.6772330
## sample estimates:
##  mean in group No mean in group Yes
##          3.007810          2.926723
```

There is no statistically significant difference between patients with and without PCOS with regards to TSH levels (p = 0.7894). This further supports the decision to not use this variable in the predictive models.

```
# Visualize the respiratory rate between patients with and without PCOS
ggplot(data, aes(pcos, rr)) + geom_boxplot(width = 0.5)
```



6.1.6 Respiratory rate

```r
# Mean and standard deviation by PCOS diagnosis for respiratory rate
data %>% dplyr::select(rr, pcos) %>% group_by(pcos) %>%
  summarise(n = n(),
            mean = mean(rr, na.rm = TRUE),
            sd = sd(rr, na.rm = TRUE))
```

```
## # A tibble: 2 x 4
##   pcos      n  mean    sd
##   <fct> <int> <dbl> <dbl>
## 1 No      364  19.2  1.71
## 2 Yes     177  19.3  1.65
```

```r
# Mean and standard deviation overall for respiratory rate
data %>% dplyr::select(rr) %>%
  summarise(n = n(),
            mean = mean(rr, na.rm = TRUE),
            sd = sd(rr, na.rm = TRUE))
```

```
## # A tibble: 1 x 3
##       n  mean    sd
##   <int> <dbl> <dbl>
## 1   541  19.2  1.69
```

```
# Independent t-test to determine if there is a difference in respiratory rates between PCOS negative a
t.test(rr ~ pcos, data = data)
```

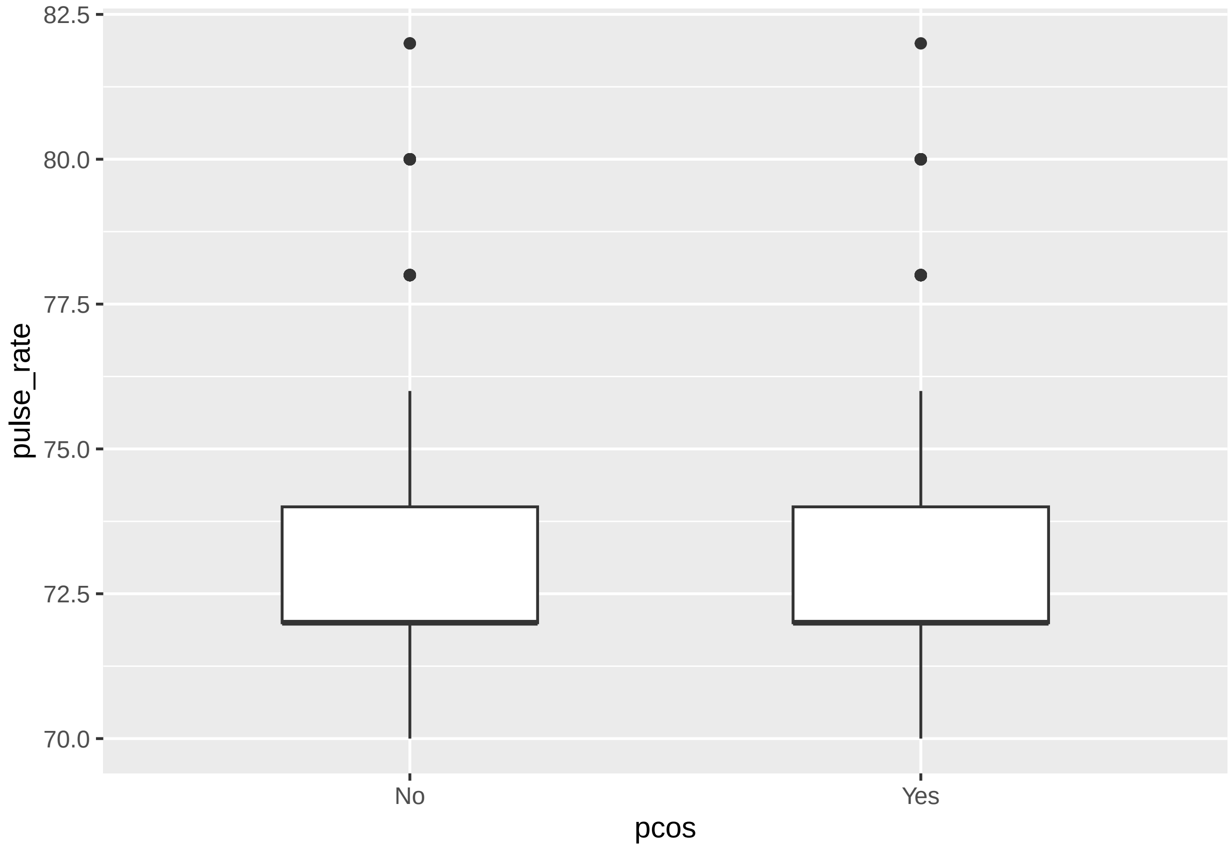
```
##
##  Welch Two Sample t-test
##
## data:  rr by pcos
## t = -0.86901, df = 360.66, p-value = 0.3854
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -0.4332740  0.1677062
## sample estimates:
##  mean in group No mean in group Yes
##          19.20055          19.33333
```

There is no statistically significant difference between patients with and without PCOS with regards to respiratory rates (p = 0.3854). This further supports the decision to not use this variable in the predictive models.

```


# Visualize the pulse rate between patients with and without PCOS
ggplot(data, aes(pcos, pulse_rate)) + geom_boxplot(width = 0.5)
```

6.1.7 Pulse rate

```
## Warning: Removed 2 rows containing non-finite values (`stat_boxplot()`).
```



```
# Mean and standard deviation by PCOS diagnosis for pulse rate
data %>% dplyr::select(pulse_rate, pcos) %>% group_by(pcos) %>%
  summarise(n = n(),
            mean = mean(pulse_rate, na.rm = TRUE),
            sd = sd(pulse_rate, na.rm = TRUE))
```

```
## # A tibble: 2 x 4
##   pcos      n  mean    sd
##   <fct> <int> <dbl> <dbl>
## 1 No      364  73.3  2.65
## 2 Yes     177  73.8  2.73
```

```
# Mean and standard deviation overall for pulse rate
data %>% dplyr::select(pulse_rate) %>%
  summarise(n = n(),
            mean = mean(pulse_rate, na.rm = TRUE),
            sd = sd(pulse_rate, na.rm = TRUE))
```

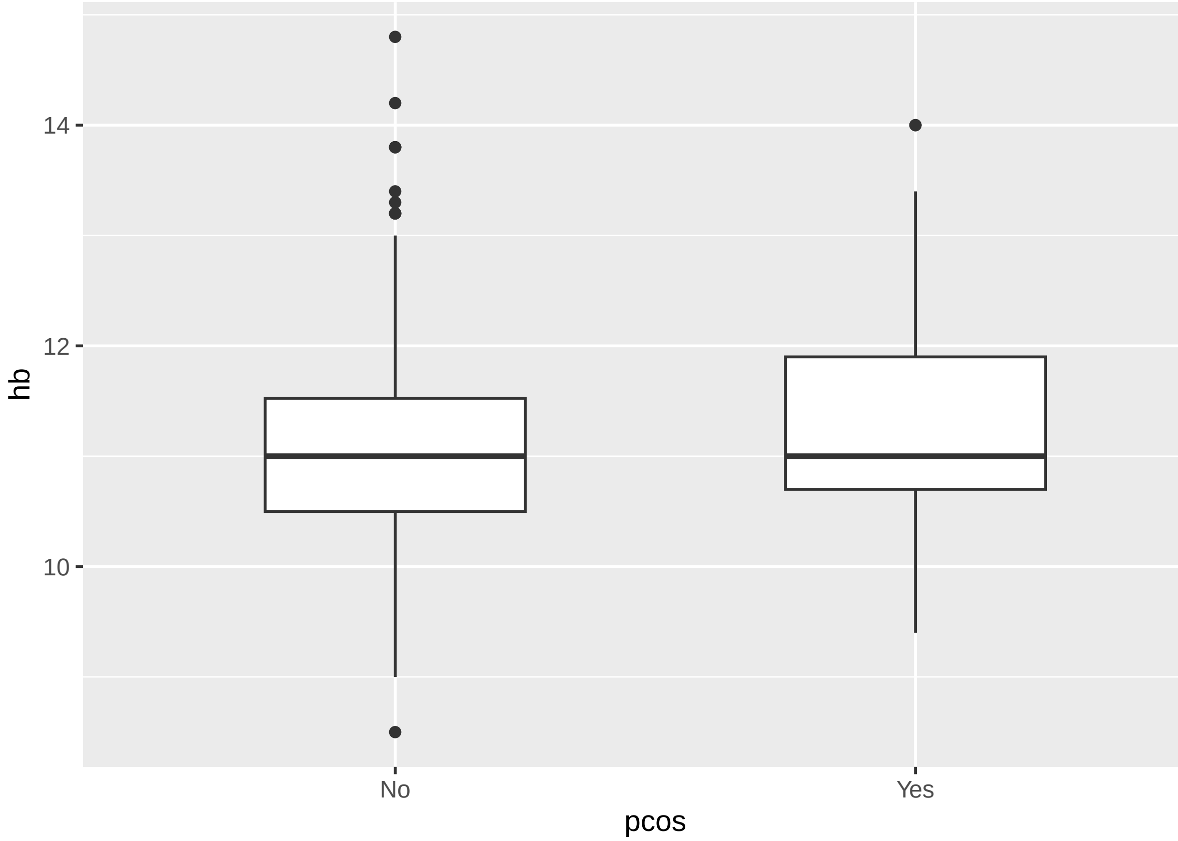
```
## # A tibble: 1 x 3
##       n  mean    sd
##   <int> <dbl> <dbl>
## 1   541  73.5  2.69
```

```
# Independent t-test to determine if there is a difference in pulse rates between PCOS negative and pos
t.test(pulse_rate ~ pcos, data = data)
```

```
##
##  Welch Two Sample t-test
##
## data:  pulse_rate by pcos
## t = -2.2108, df = 340.66, p-value = 0.02771
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -1.03695253 -0.06052851
## sample estimates:
##  mean in group No mean in group Yes
##          73.28177          73.83051
```

There is a statistically significant difference between patients with and without PCOS with regards to pulse rate (p = 0.02771). This does not support the decision to not use this variable in the predictive models. However, in view of the lack of evidence in the scientific literature for including this variable, it will not be included in the model.

```
# Visualize hemoglobin between patients with and without PCOS
ggplot(data, aes(pcos, hb)) + geom_boxplot(width = 0.5)
```



6.1.8 Hemoglobin

```r
# Mean and standard deviation by PCOS diagnosis for hemoglobin
data %>% dplyr::select(hb, pcos) %>% group_by(pcos) %>%
  summarise(n = n(),
            mean = mean(hb, na.rm = TRUE),
            sd = sd(hb, na.rm = TRUE))
```

```
## # A tibble: 2 x 4
##   pcos      n  mean    sd
##   <fct> <int> <dbl> <dbl>
## 1 No      364  11.1 0.880
## 2 Yes     177  11.3 0.831
```

```r
# Mean and standard deviation overall for hemoglobin
data %>% dplyr::select(hb) %>%
  summarise(n = n(),
            mean = mean(hb, na.rm = TRUE),
            sd = sd(hb, na.rm = TRUE))
```

```
## # A tibble: 1 x 3
##       n  mean    sd
##   <int> <dbl> <dbl>
## 1   541  11.2 0.867
```

```r
# Independent t-test to determine if there is a difference in hemoglobin between PCOS negative and posi
t.test(hb ~ pcos, data = data)
```

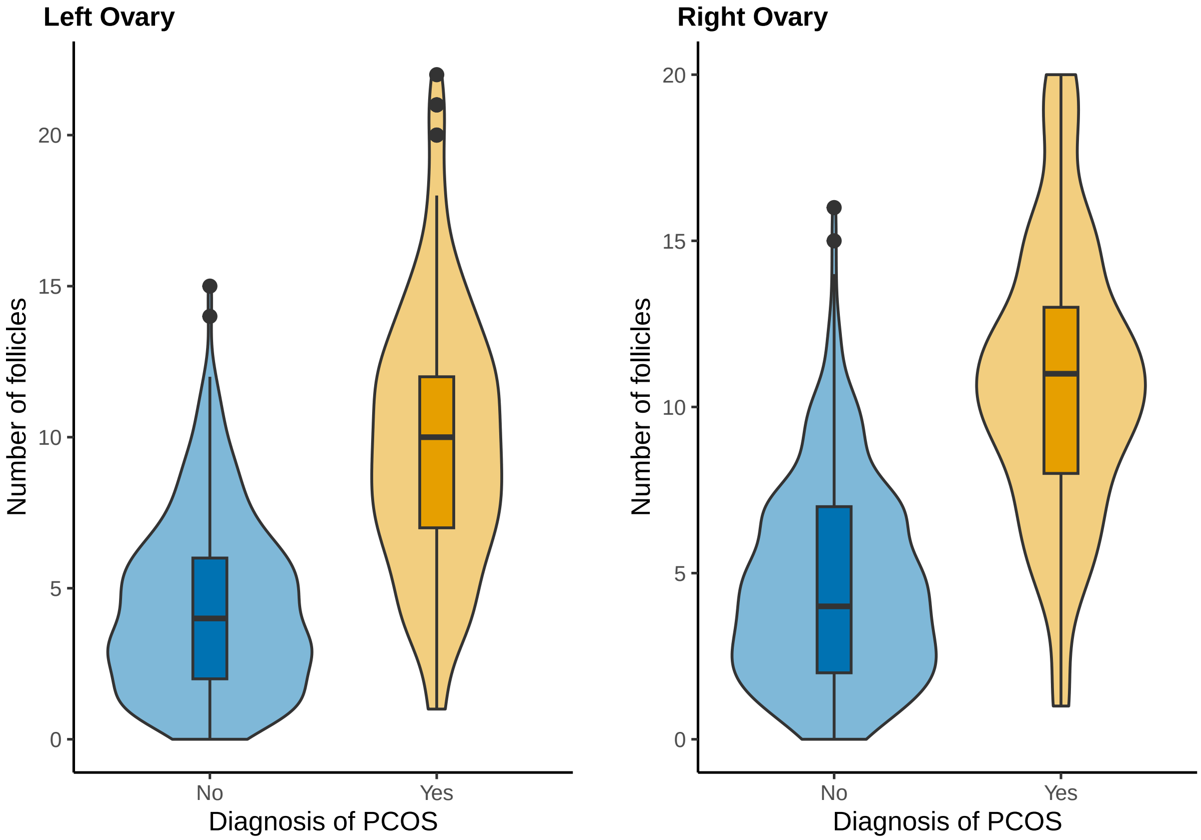
```
##
##  Welch Two Sample t-test
##
## data:  hb by pcos
## t = -2.0728, df = 367.64, p-value = 0.03888
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -0.313569336 -0.008260613
## sample estimates:
##  mean in group No mean in group Yes
##          11.10739          11.26831
```

There is a statistically significant difference between patients with and without PCOS with regards to hemoglobin (p = 0.03888). This does not support the decision to not use this variable in the predictive models. However, in view of the lack of evidence in the scientific literature for including this variable, it will not be included in the model.

6.2 Variables most correlated with PCOS

The two variables most correlated with PCOS according to the correlation coefficients in figure 5.1 are the number of follicles in the left and right ovary. The association of these two variables with a diagnosis of PCOS will be explored in further detail below through visualization and statistical tests.

```
violin_left <- ggplot(data, aes(x=pcos, y=follicle_no_l)) +
  geom_violin(aes(fill=pcos), alpha=0.5) +
  geom_boxplot(aes(fill=pcos), outlier.size=2, width=0.15) +
  scale_fill_manual(values=c("#0072B2", "#E69F00")) +
  scale_x_discrete(name = "Diagnosis of PCOS") +
  scale_y_continuous(name = "Number of follicles") +
  guides(fill="none") +
  theme_classic(10)

violin_right <- ggplot(data, aes(x=pcos, y=follicle_no_r)) +
  geom_violin(aes(fill=pcos), alpha=0.5) +
  geom_boxplot(aes(fill=pcos), outlier.size=2, width=0.15) +
  scale_fill_manual(values=c("#0072B2", "#E69F00")) +
  scale_x_discrete(name = "Diagnosis of PCOS") +
  scale_y_continuous(name = "Number of follicles") +
  guides(fill="none") +
  theme_classic(10)

figure6.2 <- plot_grid(violin_left, violin_right, labels = c('Left Ovary', 'Right Ovary'), label_size =

print(figure6.2)
```



There is clearly a large difference in the average number follicles in each ovary between women with and without a diagnosis of PCOS.This association is studied with a statistical significance test below.

```r
# Mean and standard deviation by PCOS diagnosis for number of follicles in the left
data %>% dplyr::select(follicle_no_l, pcos) %>% group_by(pcos) %>%
  summarise(n = n(),
            mean = mean(follicle_no_l, na.rm = TRUE),
            sd = sd(follicle_no_l, na.rm = TRUE))
```

```
## # A tibble: 2 x 4
##   pcos      n  mean    sd
##   <fct> <int> <dbl> <dbl>
## 1 No      364  4.35  2.81
## 2 Yes     177  9.79  4.31
```

```r
# Independent t-test to determine if there is a difference in the number of follicles in the left betwe
t.test(follicle_no_l ~ pcos, data = data)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  follicle_no_l by pcos
## t = -15.268, df = 251.34, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -6.134571 -4.732754
## sample estimates:
##  mean in group No mean in group Yes
##          4.351648          9.785311
```

```r
# Mean and standard deviation by PCOS diagnosis for number of follicles in the right
data %>% dplyr::select(follicle_no_r, pcos) %>% group_by(pcos) %>%
  summarise(n = n(),
            mean = mean(follicle_no_r, na.rm = TRUE),
            sd = sd(follicle_no_r, na.rm = TRUE))
```

```
## # A tibble: 2 x 4
##   pcos      n  mean    sd
##   <fct> <int> <dbl> <dbl>
## 1 No      364  4.64  2.93
## 2 Yes     177 10.8   4.17
```

```r
# Independent t-test to determine if there is a difference in the number of follicles in the left betwe
t.test(follicle_no_r ~ pcos, data = data)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  follicle_no_r by pcos
## t = -17.569, df = 263.24, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -6.811854 -5.438844
## sample estimates:
##  mean in group No mean in group Yes
##          4.637363         10.762712
```

Save clean object

Finally, I will save the object for future stages of this project.

```r
save(data,file =  here("data.Rdata"))
skim(data)
```

Table 11: Data summary

Name	data
Number of rows	541
Number of columns	43
Column type frequency:	
character	1
factor	10
numeric	32
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
id	0	1	1	3	0	541	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
pcos	0	1	FALSE	2	No: 364, Yes: 177
blood_group	0	1	FALSE	8	O+: 206, B+: 135, A+: 108, AB+: 42
pregnant	0	1	FALSE	2	No: 335, Yes: 206
weight_gain	0	1	FALSE	2	No: 337, Yes: 204
hair_growth	0	1	FALSE	2	No: 393, Yes: 148
skin_darkening	0	1	FALSE	2	No: 375, Yes: 166
hair_loss	0	1	FALSE	2	No: 296, Yes: 245
pimples	0	1	FALSE	2	No: 276, Yes: 265
fast_food	1	1	FALSE	2	Yes: 278, No: 262
reg_exercise	0	1	FALSE	2	No: 407, Yes: 134

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	31.43	5.41	20.00	28.00	31.00	35.00	48.00	
weight	0	1	59.64	11.03	31.00	52.00	59.00	65.00	108.00	
height	0	1	156.48	6.03	137.00	152.00	156.00	160.00	180.00	
bmi	0	1	24.31	4.06	12.42	21.64	24.24	26.63	38.90	
pulse_rate	2	1	73.46	2.69	70.00	72.00	72.00	74.00	82.00	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
rr	0	1	19.24	1.69	16.00	18.00	18.00	20.00	28.00	
hb	0	1	11.16	0.87	8.50	10.50	11.00	11.70	14.80	
cycle	0	1	2.56	0.90	2.00	2.00	2.00	4.00	5.00	
cycle_length	0	1	4.94	1.49	0.00	4.00	5.00	5.00	12.00	
marriage_status	1	1	7.68	4.80	0.00	4.00	7.00	10.00	30.00	
no_of_abortions	0	1	0.29	0.69	0.00	0.00	0.00	0.00	5.00	
i_betahcg	0	1	664.55	3348.92	1.30	1.99	20.00	297.21	32460.97	
ii_betahcg	0	1	238.23	1603.83	0.99	1.99	1.99	97.63	25000.00	
fsh	1	1	0.65	0.23	-0.68	0.52	0.69	0.81	1.82	
lh	1	1	0.26	0.45	-1.70	0.01	0.36	0.56	1.17	
fsh_lh_ratio	2	1	0.39	0.38	-0.64	0.15	0.34	0.60	2.51	
hip	0	1	37.99	3.97	26.00	36.00	38.00	40.00	48.00	
waist	0	1	33.84	3.60	24.00	32.00	34.00	36.00	47.00	
waist_hip_ratio	0	1	0.89	0.05	0.76	0.86	0.89	0.93	0.98	
tsh	0	1	2.98	3.76	0.04	1.48	2.26	3.57	65.00	
amh	2	1	0.56	0.43	-1.00	0.30	0.57	0.84	1.51	
prl	0	1	24.32	14.97	0.40	14.52	21.92	29.89	128.24	
vitd3	2	1	28.89	12.54	0.00	20.75	25.90	34.30	90.00	
prg	0	1	-0.44	0.24	-1.33	-0.60	-0.49	-0.35	1.93	
rbs	0	1	1.99	0.07	1.78	1.96	2.00	2.03	2.54	
bp_systolic	1	1	114.85	5.92	100.00	110.00	110.00	120.00	140.00	
bp_diastolic	1	1	77.06	4.72	60.00	70.00	80.00	80.00	100.00	
follicle_no_l	0	1	6.13	4.23	0.00	3.00	5.00	9.00	22.00	
follicle_no_r	0	1	6.64	4.44	0.00	3.00	6.00	10.00	20.00	
avg_f_size_l	0	1	15.02	3.57	0.00	13.00	15.00	18.00	24.00	
avg_f_size_r	0	1	15.45	3.32	0.00	13.00	16.00	18.00	24.00	
endometrium	0	1	8.48	2.17	0.00	7.00	8.50	9.80	18.00	

Conclusion

This EDA was very helpful to familiarize myself with the data, clean it, and identify any pattern that could potentially need to be addressed in the future of my analysis. I was able to flag individuals with missing observations, remove outliers, transform some variables so that they had a higher variability range, and observe the correlation between my variables.

Session info

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur ... 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
```

```
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] gmodels_2.18.1.1   epiDisplay_3.5.0.2 nnet_7.3-18        MASS_7.3-58.1
##  [5] survival_3.4-0     foreign_0.8-83     cowplot_1.1.1      skimr_2.1.5
##  [9] knitr_1.41         DataExplorer_0.8.2 janitor_2.1.0      readxl_1.4.1
## [13] here_1.0.1         forcats_0.5.2      stringr_1.4.1      dplyr_1.0.10
## [17] purrr_0.3.5        readr_2.1.3        tidyr_1.2.1        tibble_3.1.8
## [21] ggplot2_3.4.0      tidyverse_1.3.2
##
## loaded via a namespace (and not attached):
##  [1] fs_1.5.2            lubridate_1.9.0     httr_1.4.4
##  [4] rprojroot_2.0.3     repr_1.1.4          tools_4.2.2
##  [7] backports_1.4.1     utf8_1.2.2          R6_2.5.1
## [10] DBI_1.1.3           colorspace_2.0-3    withr_2.5.0
## [13] tidyselect_1.2.0    gridExtra_2.3       compiler_4.2.2
## [16] cli_3.4.1           rvest_1.0.3         xml2_1.3.3
## [19] labeling_0.4.2      scales_1.2.1        digest_0.6.30
## [22] rmarkdown_2.18      base64enc_0.1-3     pkgconfig_2.0.3
## [25] htmltools_0.5.3     dbplyr_2.2.1        fastmap_1.1.0
## [28] highr_0.9           htmlwidgets_1.5.4   rlang_1.0.6
## [31] rstudioapi_0.14     farver_2.1.1        generics_0.1.3
## [34] jsonlite_1.8.3      gtools_3.9.4        googlesheets4_1.0.1
## [37] magrittr_2.0.3      Matrix_1.5-1        Rcpp_1.0.9
## [40] munsell_0.5.0       fansi_1.0.3         lifecycle_1.0.3
## [43] stringi_1.7.8       yaml_2.3.6          snakecase_0.11.0
## [46] plyr_1.8.8          grid_4.2.2          gdata_2.18.0.1
## [49] parallel_4.2.2      crayon_1.5.2        lattice_0.20-45
## [52] haven_2.5.1         splines_4.2.2       hms_1.1.2
## [55] pillar_1.8.1        igraph_1.3.5        reshape2_1.4.4
## [58] reprex_2.0.2        glue_1.6.2          evaluate_0.18
## [61] data.table_1.14.6   modelr_0.1.10       vctrs_0.5.1
## [64] tzdb_0.3.0          networkD3_0.4       cellranger_1.1.0
## [67] gtable_0.3.1        assertthat_0.2.1    xfun_0.35
## [70] broom_1.0.1         googledrive_2.0.0   gargle_1.2.1
## [73] timechange_0.1.1    ellipsis_0.3.2
```