MEDI 504B - Exploratory Data Analysis

Erick Navarro

2023-01-11

Contents

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.

Load the data

First, I will load the packages that I will use throughout the analysis and the data.

```
library(tidyverse)
library(here)
library(readxl)
library(janitor)
library(DataExplorer)
library(knitr)

data = read_excel(here("PCOS_data_without_infertility.xlsx"), sheet = "Full_new") %>%
  clean_names()
```

Clean the data

Then, I will take a look at the data, remove redundant variables, and make sure that each column is formatted in the right data type.

```
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~
```

By taking a quick look at the data, I can observe that sl_no and patient_file have apparently the same information. Also, the last column (x45) seems to be empty. I will remove those variables and change the format of the variables when needed.

```
#looks like sl_no and patient_file_no are the same column?
all(identical(data$sl_no, data$patient_file_no))
```

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

```
#Also, column x45 looks like empty. I checked the excel file and it is an empty column

#Clean and change the format of the dataset
data = data %>%
  select(-c(sl_no, x45)) %>%
  mutate(patient_file_no = as.character(patient_file_no),
         pcos_y_n = as.factor(pcos_y_n),
         ii_beta_hcg_m_iu_m_l = case_when(ii_beta_hcg_m_iu_m_l == "1.99."~ "1.99", #I found
                                          #this typo when exploring the missing data and
                                          # checking the excel file of said individual
                                          TRUE ~ ii_beta_hcg_m_iu_m_l),
         ii_beta_hcg_m_iu_m_l = as.numeric(ii_beta_hcg_m_iu_m_l),
         amh_ng_m_l = as.numeric(amh_ng_m_l),
         #Recode the blood group to the right variable
         blood_group = case_when(blood_group == 11 ~ "A+",
                                 blood_group == 12 ~ "A-",
                                 blood_group == 13 ~ "B+",
                                 blood_group == 14 ~ "B-",
                                 blood_group == 15 ~ "O+",
```

```
                                blood_group == 16 ~ "O-",
                                blood_group == 17 ~ "AB+",
                                blood_group == 18 ~ "AB-",),
         pregnant_y_n = as.factor(pregnant_y_n),
         weight_gain_y_n = as.factor(weight_gain_y_n),
         hair_growth_y_n = as.factor(hair_growth_y_n),
         skin_darkening_y_n = as.factor(skin_darkening_y_n),
         hair_loss_y_n = as.factor(hair_loss_y_n),
         pimples_y_n = as.factor(pimples_y_n),
         fast_food_y_n = as.factor(fast_food_y_n),
         reg_exercise_y_n = as.factor(reg_exercise_y_n))
```

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

```
glimpse(data)
```

```
## Rows: 541
## Columns: 43
## $ patient_file_no      <chr> "1", "2", "3", "4", "5", "6", "7", "8", "9", "10"~
## $ pcos_y_n             <fct> 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          <chr> "O+", "O+", "A+", "B+", "A+", "O+", "A+", "B+", "~
## $ 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         <fct> 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 <dbl> 1.99, 1.99, 494.08, 1.99, 801.45, 1.99, 1.99, 100~
## $ 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           <dbl> 2.07, 1.53, 6.63, 1.22, 2.26, 6.74, 3.05, 1.54, 1~
## $ 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      <fct> 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0~
## $ hair_growth_y_n      <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0~
## $ skin_darkening_y_n   <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0~
## $ hair_loss_y_n        <fct> 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0~
## $ pimples_y_n          <fct> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0~
```

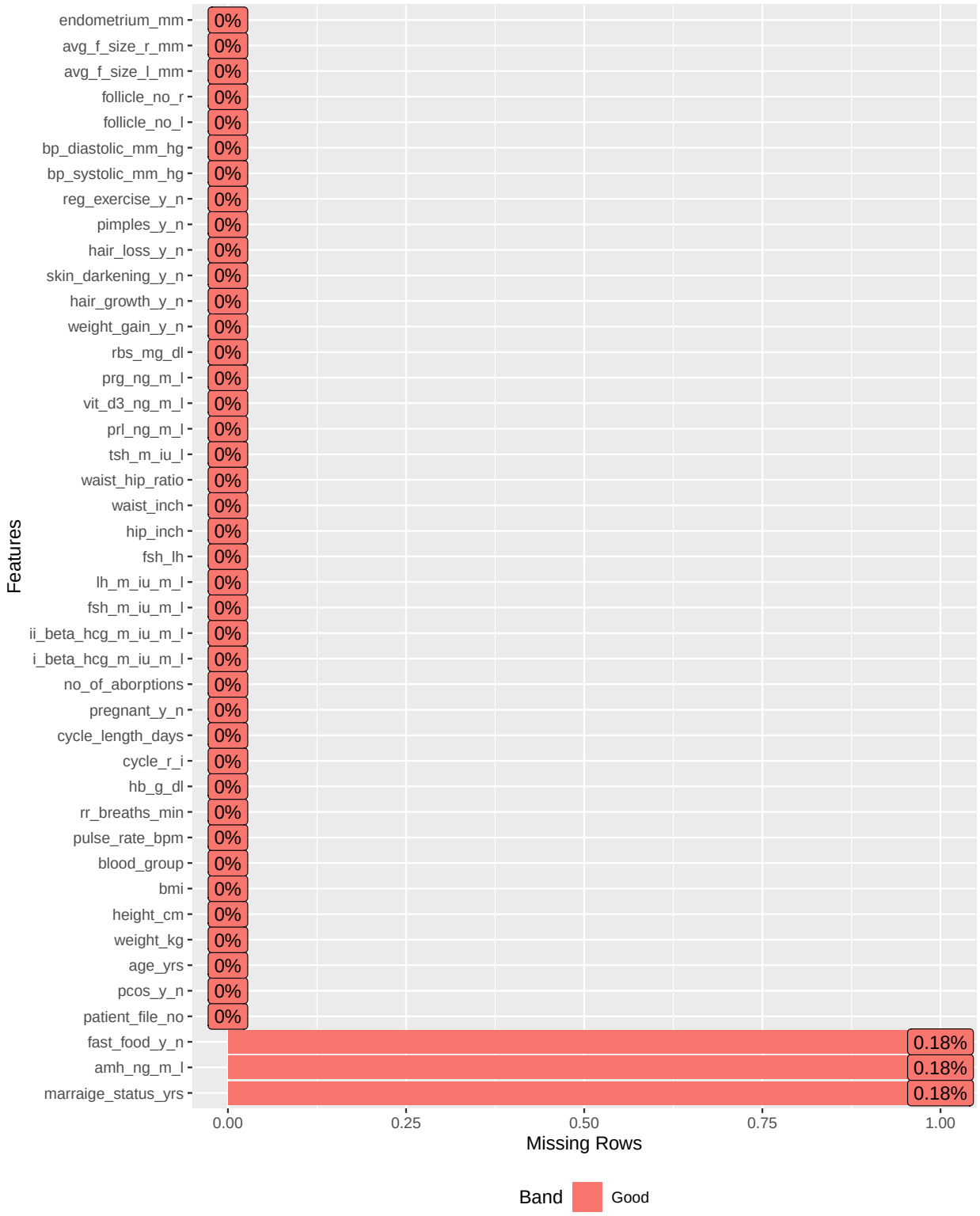
```
## $ fast_food_y_n        <fct> 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0~
## $ reg_exercise_y_n     <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0~
## $ bp_systolic_mm_hg    <dbl> 110, 120, 120, 120, 120, 110, 120, 120, 120, 110,~
## $ bp_diastolic_mm_hg   <dbl> 80, 70, 80, 70, 80, 70, 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~
```

Now the data is in the right format, so we can proceed to plot it and explore it.

Explore missing values

As a first step of the exploration, I will check the missingness of each variable in my dataset.

```
plot_missing(data)
```



We can see that there is a missing value in the variables fast_food_y_n, marriage_status_yrs and amh_ng_m_l. I will explore if it's missing in the same person

```
data %>%
  filter(is.na(fast_food_y_n) |
```

```
          is.na(marraige_status_yrs) |
          is.na(amh_ng_m_l)) %>%
  select(c(patient_file_no, fast_food_y_n, marraige_status_yrs, amh_ng_m_l)) %>%
  knitr::kable()
```

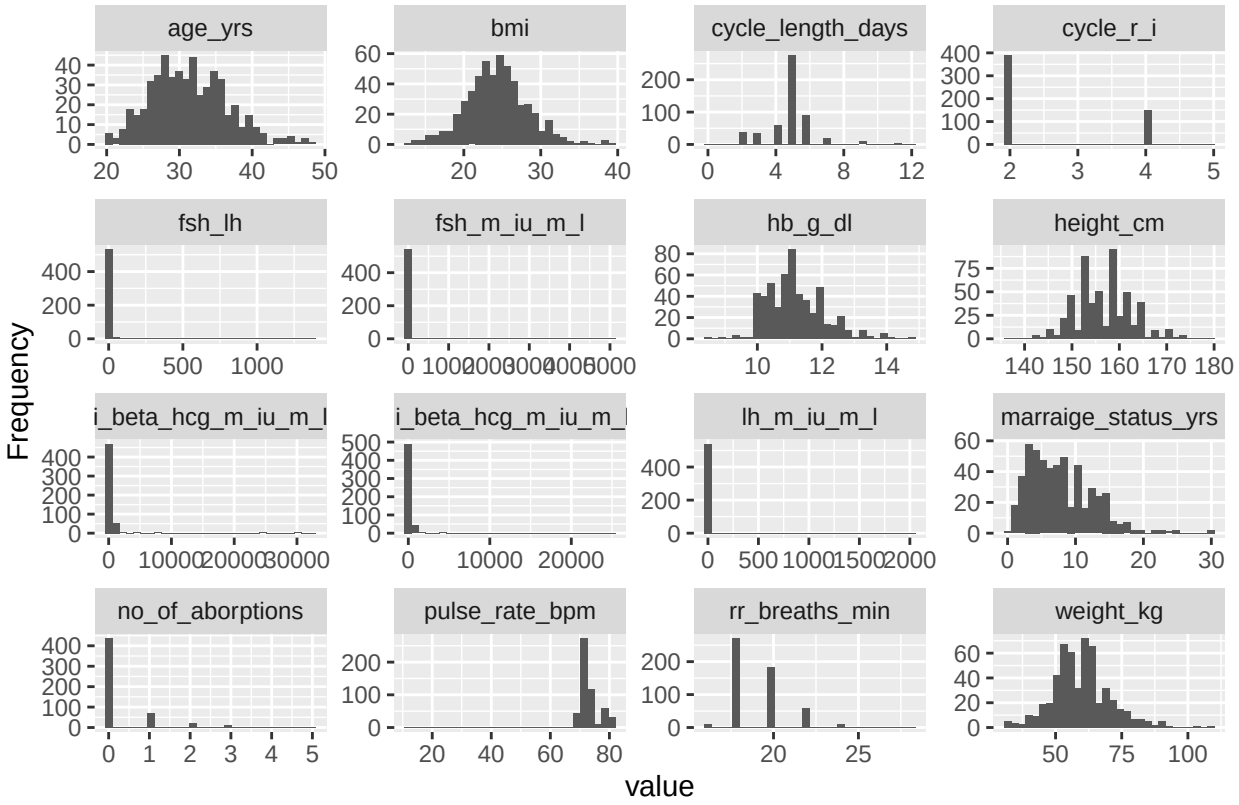
patient_file_no	fast_food_y_n	marraige_status_yrs	amh_ng_m_l
157	NA	5	5.27
306	0	9	NA
459	0	NA	6.60

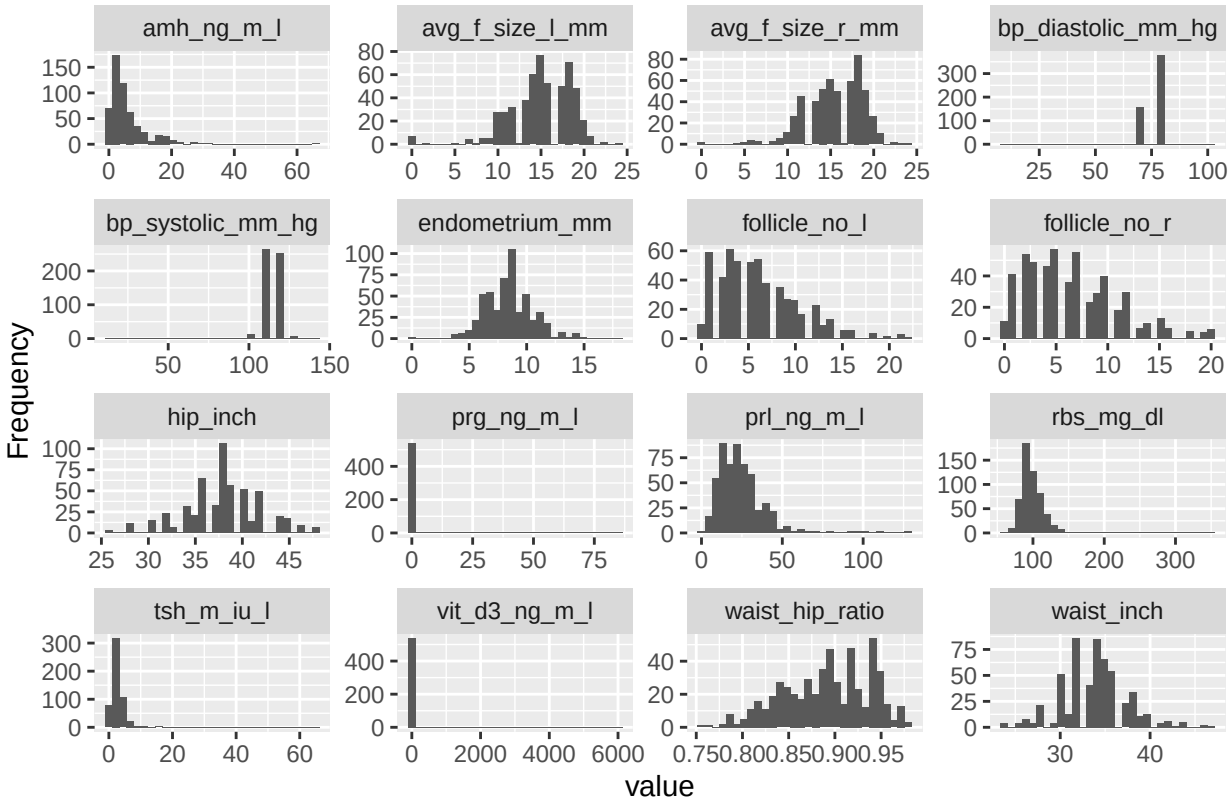
We can observe that the individuals with missing observations are different. I will flag individuals 157, 306 and 456 in case this creates a problem later in the analysis, but since each of them have only one missing variable, I don't think it's needed to remove them.

Explore variation of continuous variables

Now, I will explore the variability of the continuous variables in the dataset.

```
plot_histogram(data)
```





Explore outliers

We can observe that the variables prg_ng_m_l and vit_d3_ng_m_l, fsh_lh, fsh_m_iu_m_l, i_beta_hcg_m_iu_m_l, ii_beta_hcg_m_iu_m_l and lh_m_iu_m_l and pulse_rate_bpm 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 being non-variable in these variables. This can be checked by observing the summary of said variables. I will explore them more in detail.

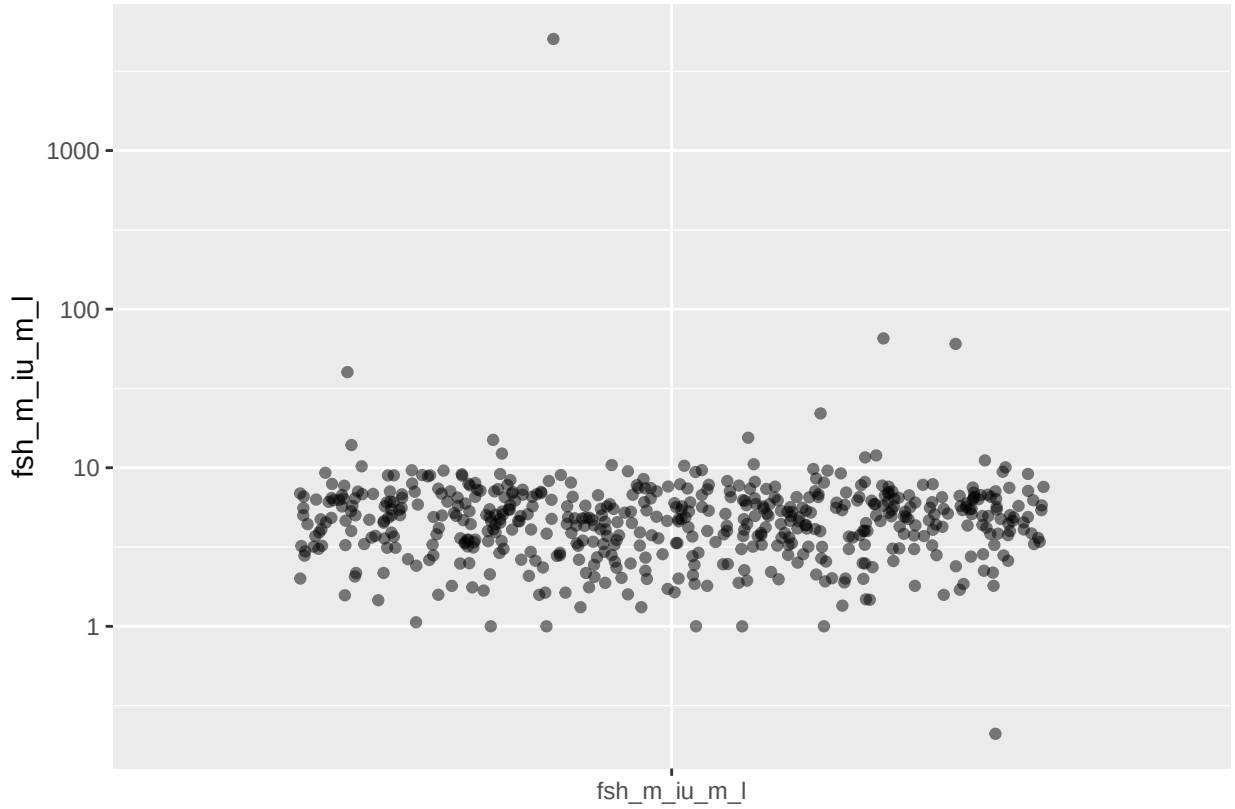
```
data %>%
  select(prg_ng_m_l, vit_d3_ng_m_l, fsh_lh, fsh_m_iu_m_l, i_beta_hcg_m_iu_m_l,
           ii_beta_hcg_m_iu_m_l, lh_m_iu_m_l, pulse_rate_bpm) %>%
  rownames_to_column(var = "ID") %>%
  pivot_longer(-ID, names_to = "variables", values_to = "data") %>%
  group_by(variables) %>%
  summarise(mean = 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_lh	6.9048308	1.416244	2.169231	3.959184	1372.826	0.0021457
fsh_m_iu_m_l	14.6018318	3.300000	4.850000	6.410000	5052.000	0.2100000
i_beta_hcg_m_iu_m_l	664.5492348	1.990000	20.000000	297.210000	32460.970	1.3000000
ii_beta_hcg_m_iu_m_l	238.2329926	1.990000	1.990000	97.630000	25000.000	0.9900000
lh_m_iu_m_l	6.4699187	1.020000	2.300000	3.680000	2018.000	0.0200000
prg_ng_m_l	0.6109445	0.250000	0.320000	0.450000	85.000	0.0470000
pulse_rate_bpm	73.2476895	72.000000	72.000000	74.000000	82.000	13.0000000
vit_d3_ng_m_l	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 like non-variable because there are outliers that drag the distributions. I will check which samples are outliers for each of these variables.

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

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



```
data %>%
  filter(fsh_m_iu_m_l > 1000) %>%
  pull(patient_file_no)
```

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

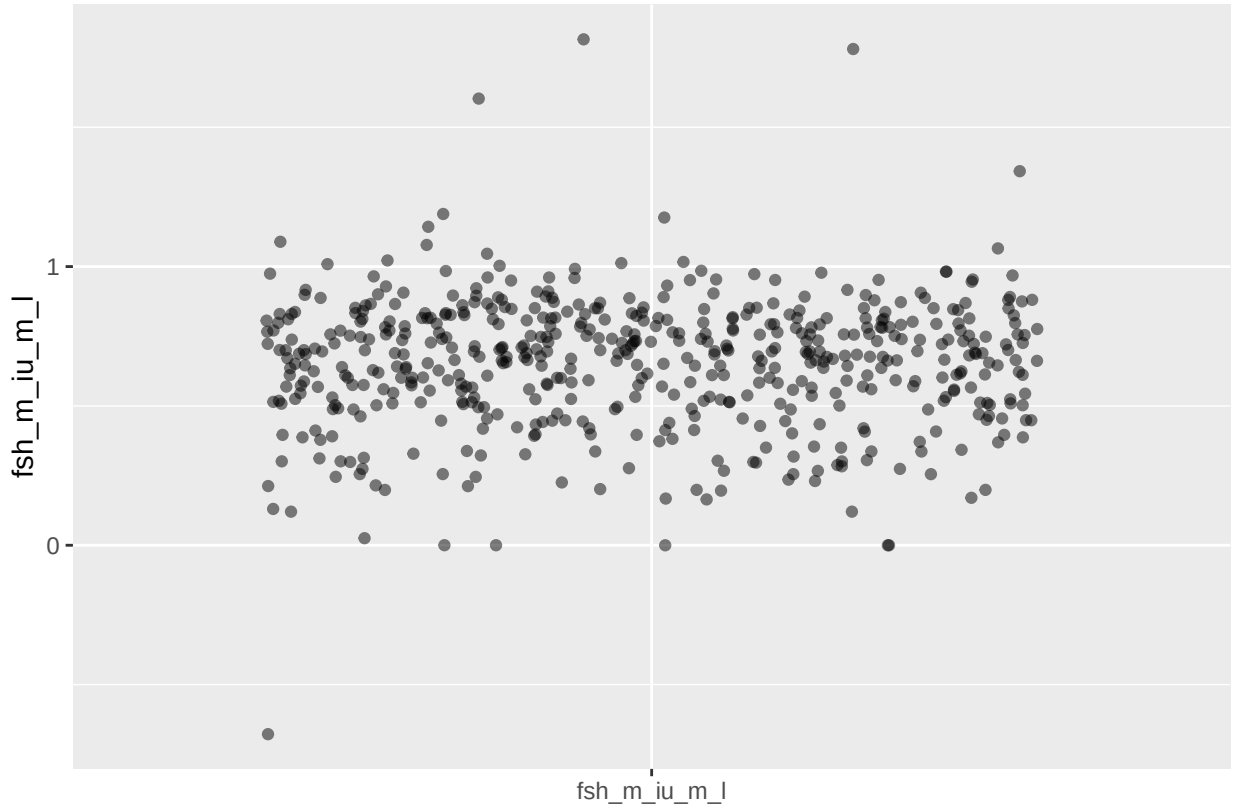
According to reference values, this sample has an impossible biological value. Therefore, I will set this observation and related variables to NA.

```
data[data$patient_file_no == 330,"fsh_m_iu_m_l" ] = NA
data[data$patient_file_no == 330,"fsh_lh" ] = NA
```

I will also log10 transform the values because looking at the quartiles above, data is compressed in the left side of the histogram

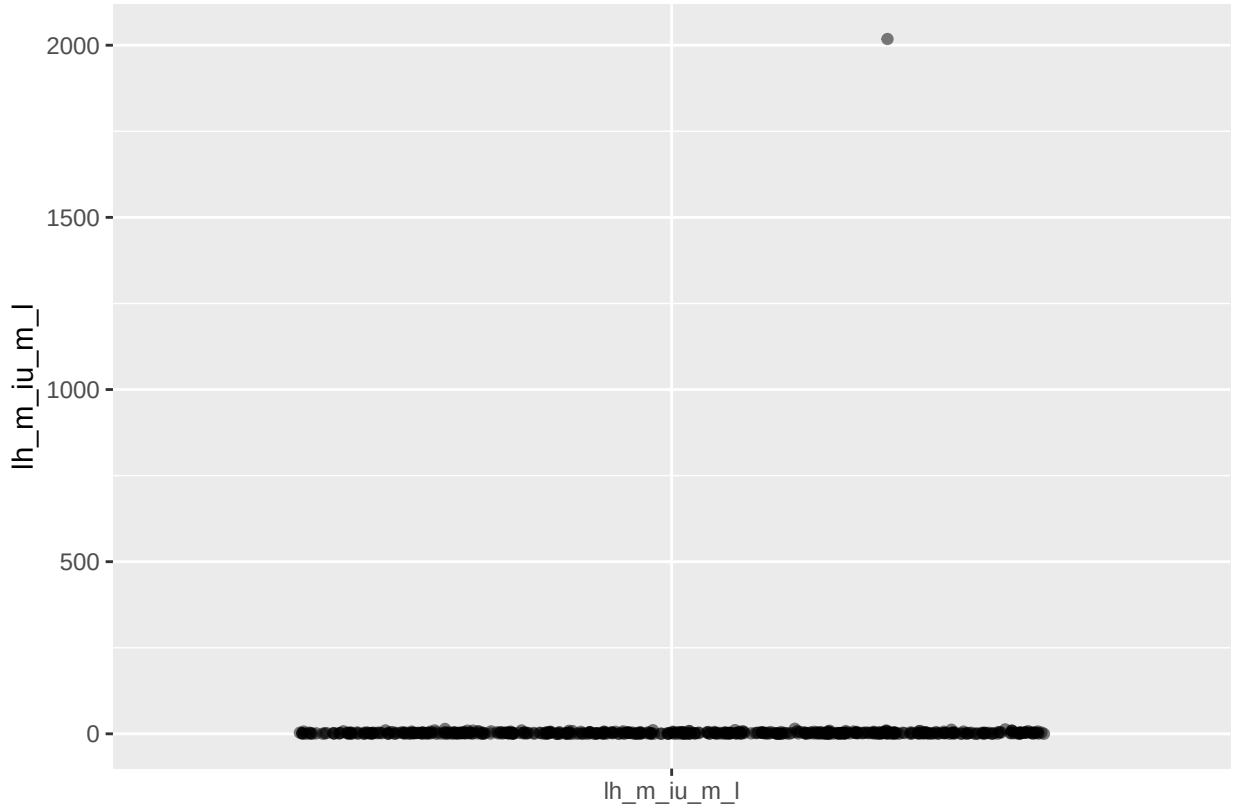
```
data = data %>%
  mutate(fsh_m_iu_m_l = log10(fsh_m_iu_m_l))

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

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



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

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




```
data %>%
  filter(lh_m_iu_m_l > 1000) %>%
  pull(patient_file_no)
```

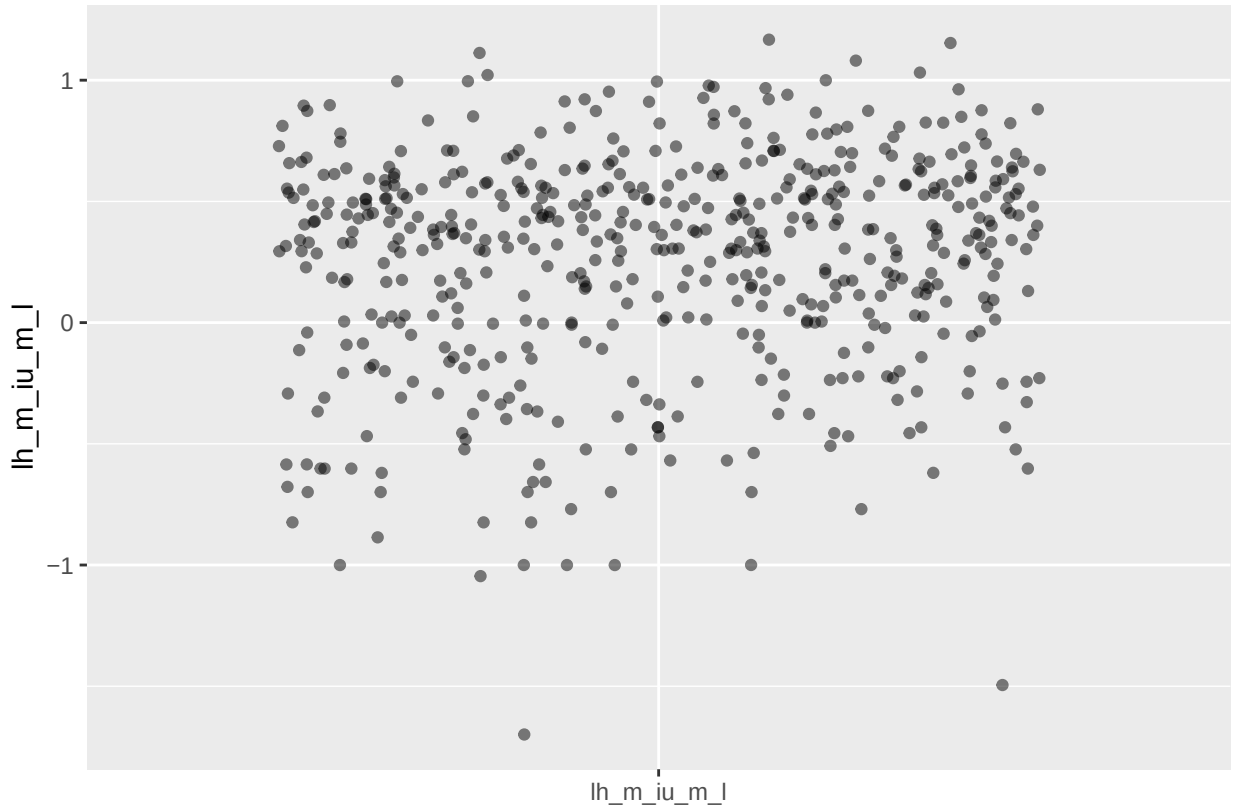
```
## [1] "456"
```

The individual 456 has a LH level outside of the reported reference levles. Therefore, I will remove this variable and related ones and also log10-transform it. 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$patient_file_no == 456,"lh_m_iu_m_l" ] = NA
data[data$patient_file_no == 456,"fsh_lh" ] = NA

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

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

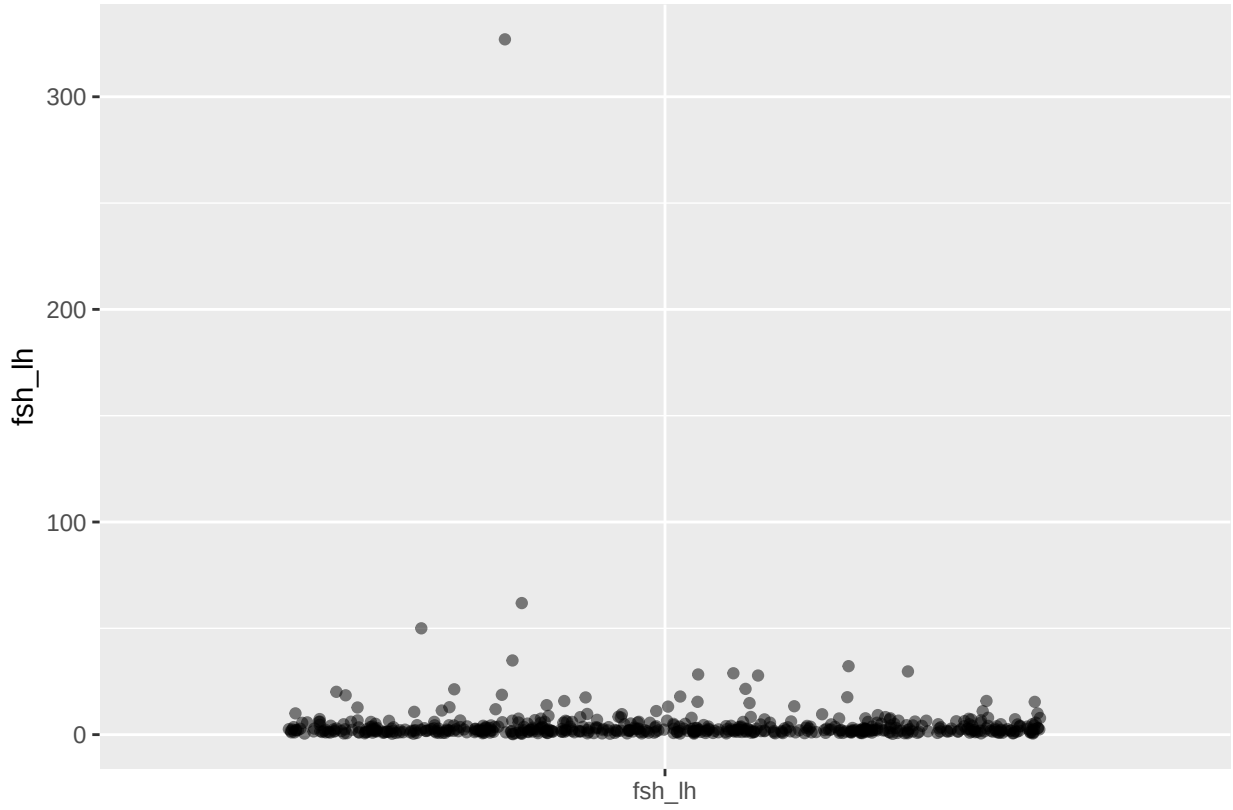
```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



FSH/LH ratio Since I have already removed outliers from the FSH and LH variables, the remaining outlier here should be biologically occurrent. Then, I will only log10-transform this variable.

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

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

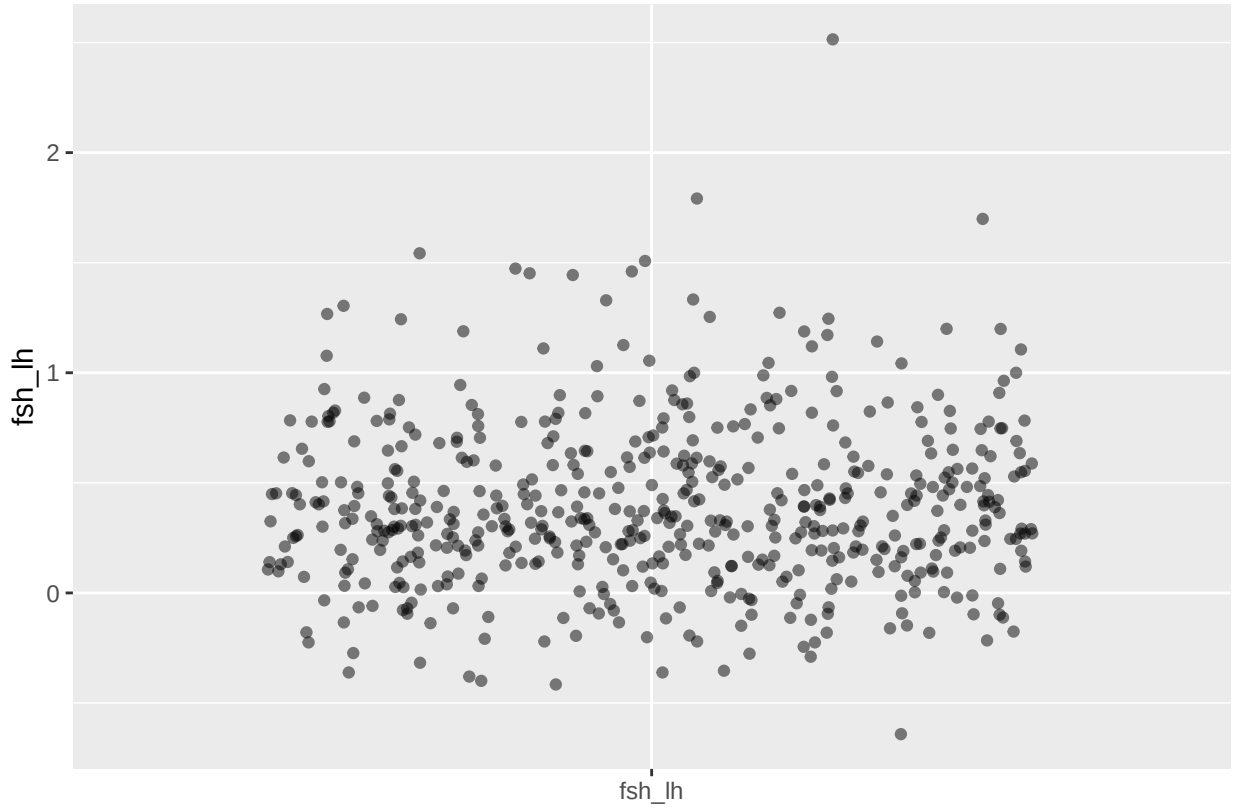



```
data %>%
  filter(fsh_lh > 250) %>%
  pull(patient_file_no)
```

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

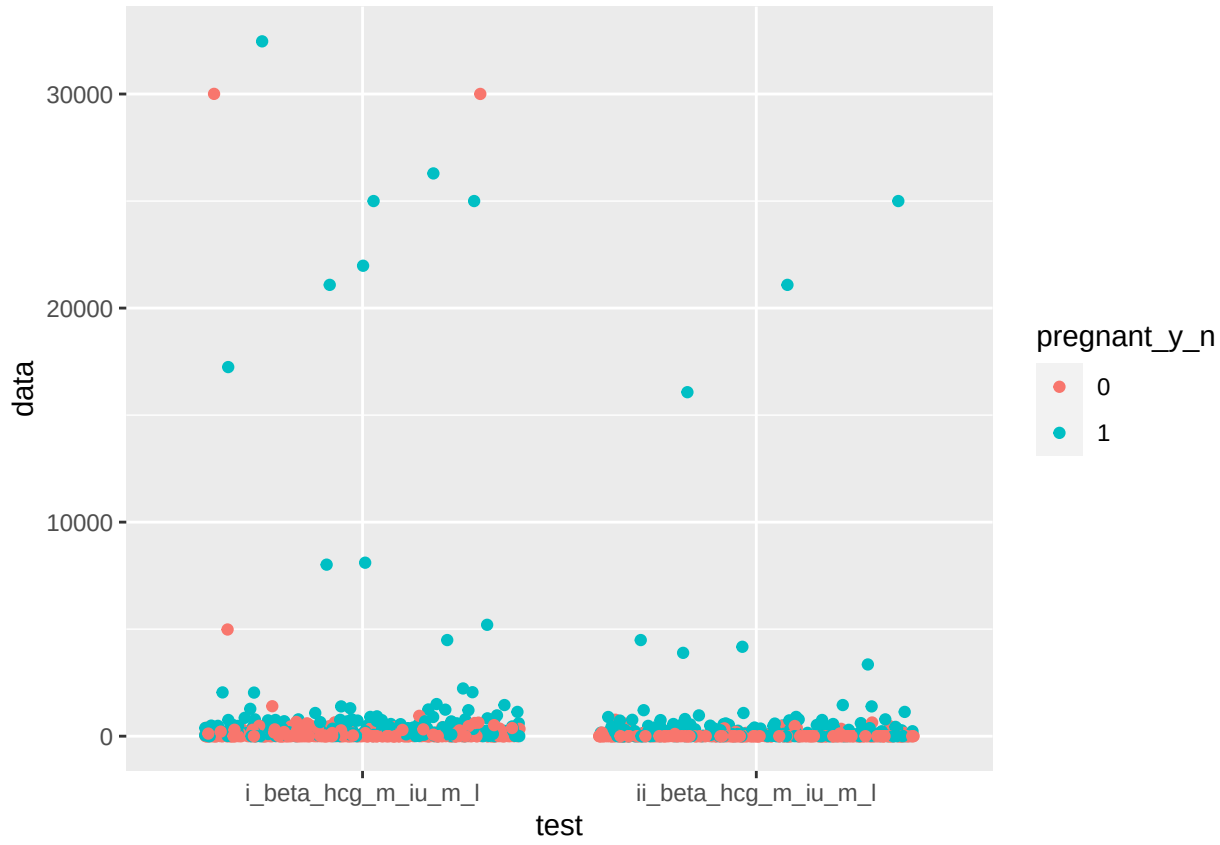
```
#I will flag this patient in case it pops out somewhere else in the analysis.

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

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

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

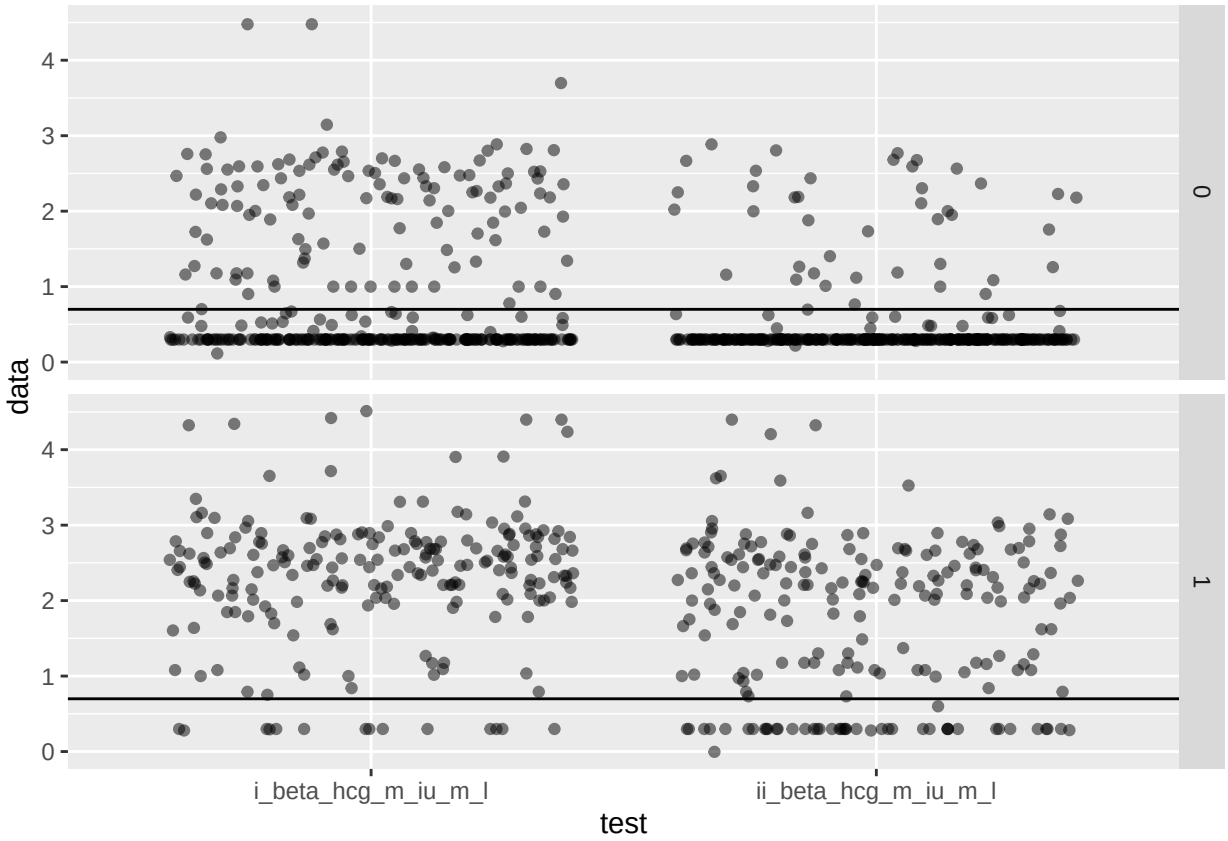


Human chorionic gonadotropin (hCG) in the blood According to reference levels, the values present in our dataset are within the expected biological range. Then, I will just log-10 transform these variables.

```
data %>%
  select(c(patient_file_no, i_beta_hcg_m_iu_m_l,ii_beta_hcg_m_iu_m_l, pregnant_y_n)) %>%
  pivot_longer(- c(patient_file_no,pregnant_y_n ), names_to = "test", values_to = "data") %>%
  ggplot(aes(x = test, y = data, color = pregnant_y_n)) +
  geom_jitter()
```



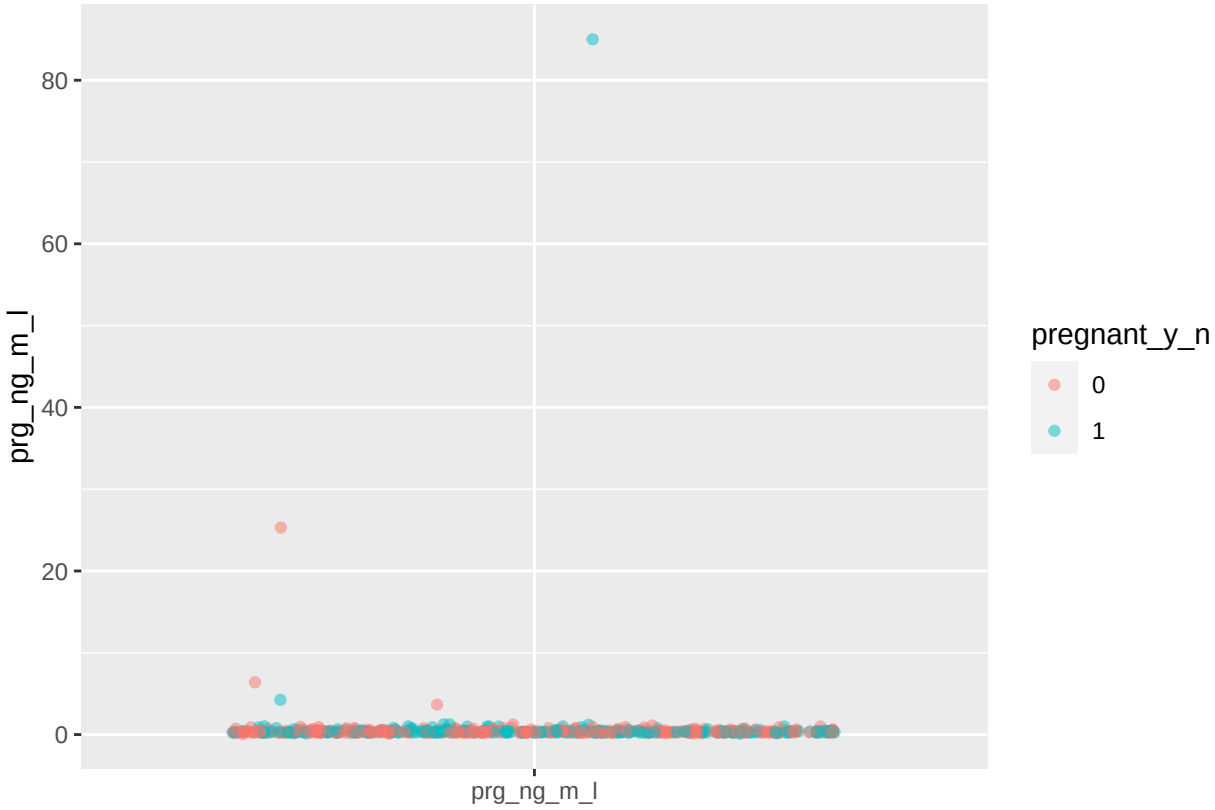

```
data = data %>%
  mutate(i_beta_hcg_m_iu_m_l = log10(i_beta_hcg_m_iu_m_l),
         ii_beta_hcg_m_iu_m_l = log10(ii_beta_hcg_m_iu_m_l))

#look at the data after transformation
data %>%
  select(c(patient_file_no, i_beta_hcg_m_iu_m_l,ii_beta_hcg_m_iu_m_l, pregnant_y_n)) %>%
  pivot_longer(- c(patient_file_no,pregnant_y_n ), 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_y_n")
```



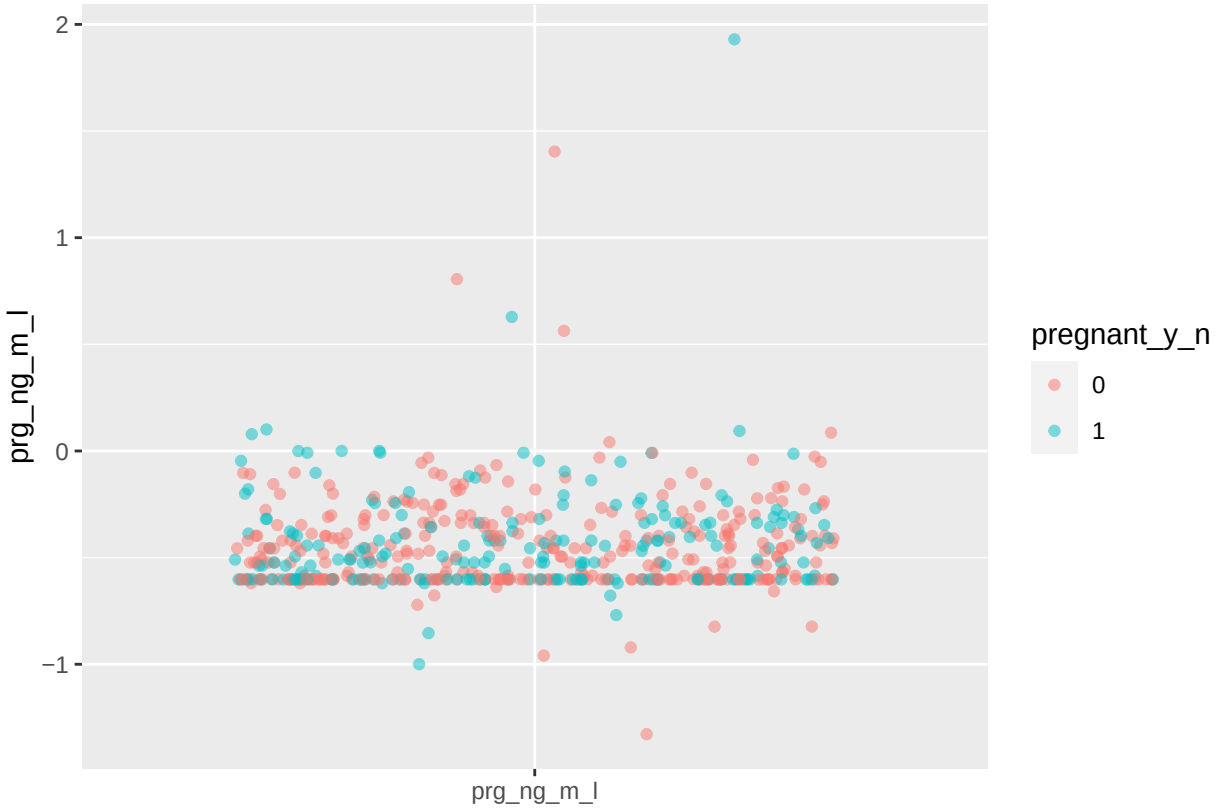
We can observe that even though both tests are supposed to measure the same hormone in blood, they do not provide similar results for many cases. Since the dataset does not provide more information about the tests, there is no way to know which one is more accurate, which would allow me to drop one of the variables. Also, it is weird that non-pregnant women are supposed to have a level of less than 5 mIU/mL. With the last plot, we can see that this is condition is not met; test 1 seems to have more false negatives, and test 2 seems to have more false positives. For now, I will keep both of these variables.

Progesterone According to reference levels, the values in the dataset seem to be biologically possible. Then, I will only log-10 transform the data.

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



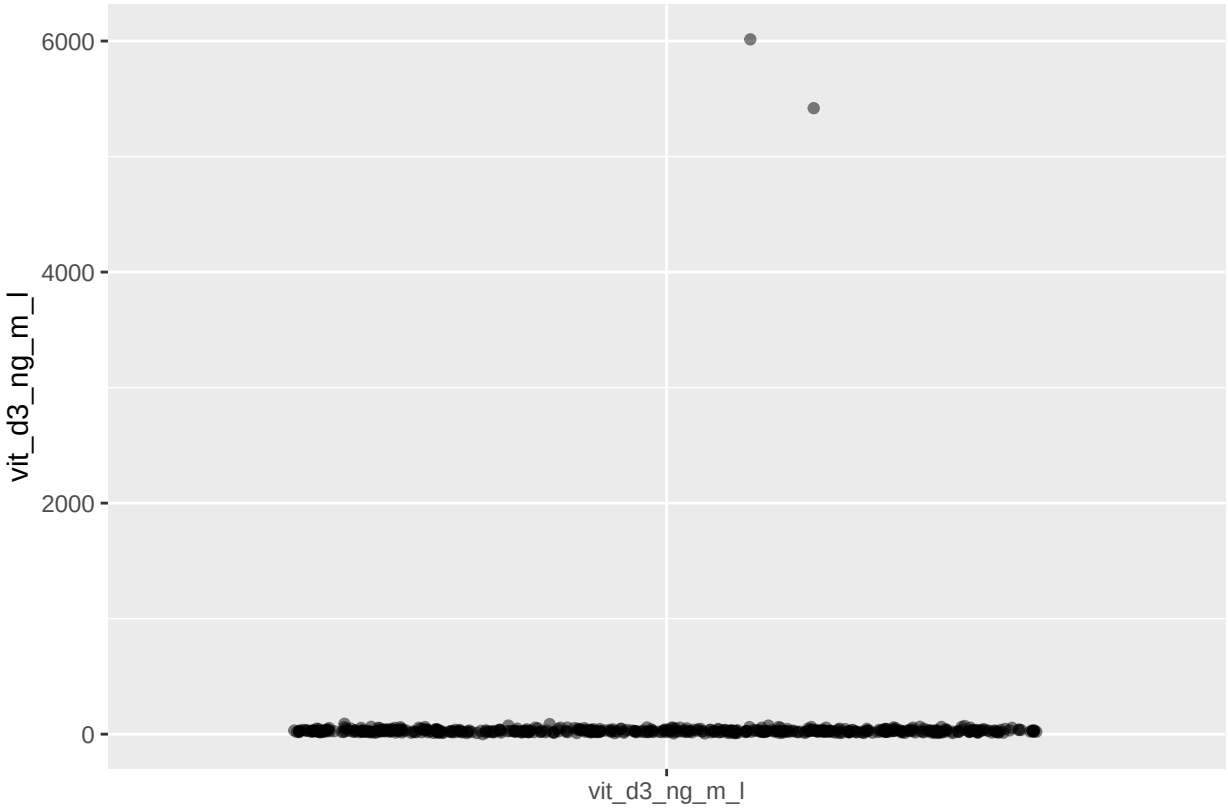

```
data = data %>%
  mutate(prg_ng_m_l = log10(prg_ng_m_l))

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



Progesterone levels are very variable depending on the mentstrual cycle stage of the person, so everything looks in order. 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 red dot in the upper part.

Vitamin D3 It has been reported that a normal range of vitamin D is 30 to 74 ng/mL, and that side effects or toxicity can occur when blood concentrations reach 88 ng/mL or greater. Then, the outliers in the data are biologically not possible.

```
data %>%
  ggplot(aes(x = "vit_d3_ng_m_l", y = vit_d3_ng_m_l)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```



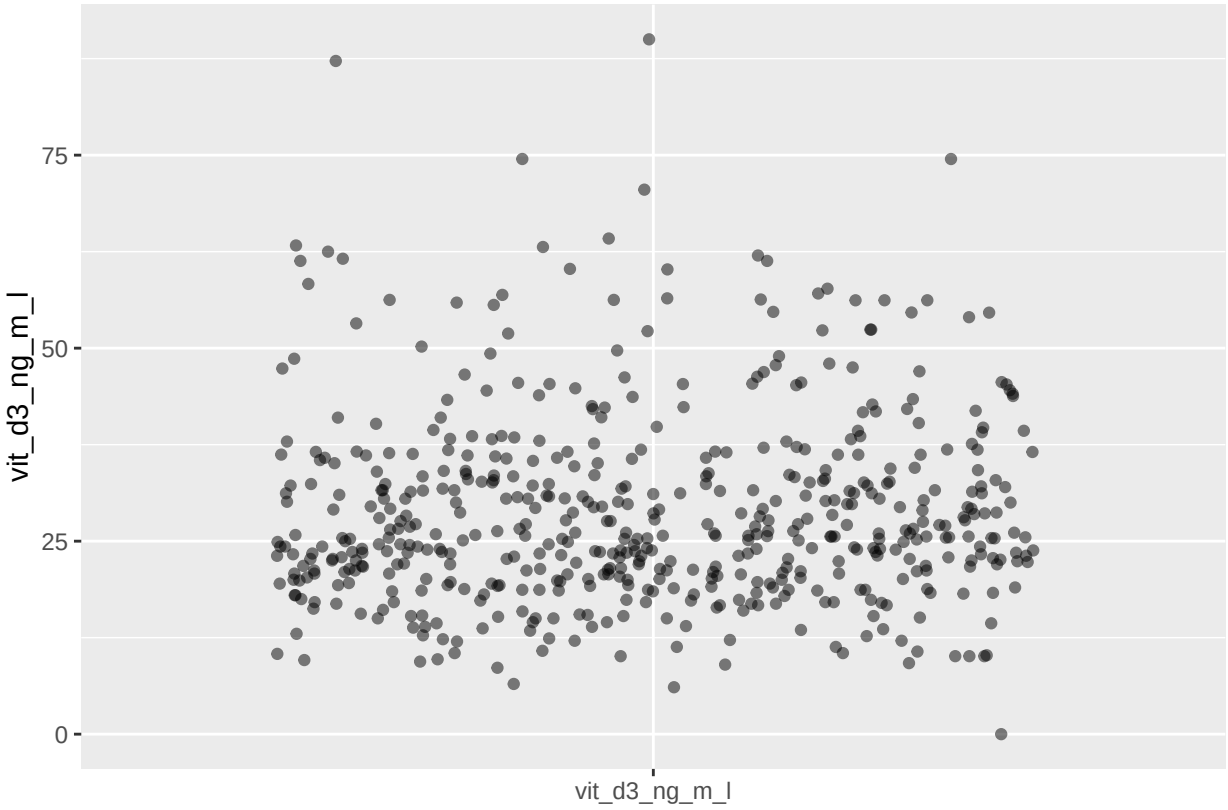
Therefore, I will remove those values.

```
data %>%
  filter(vit_d3_ng_m_l > 90) %>%
  pull(patient_file_no)
```

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

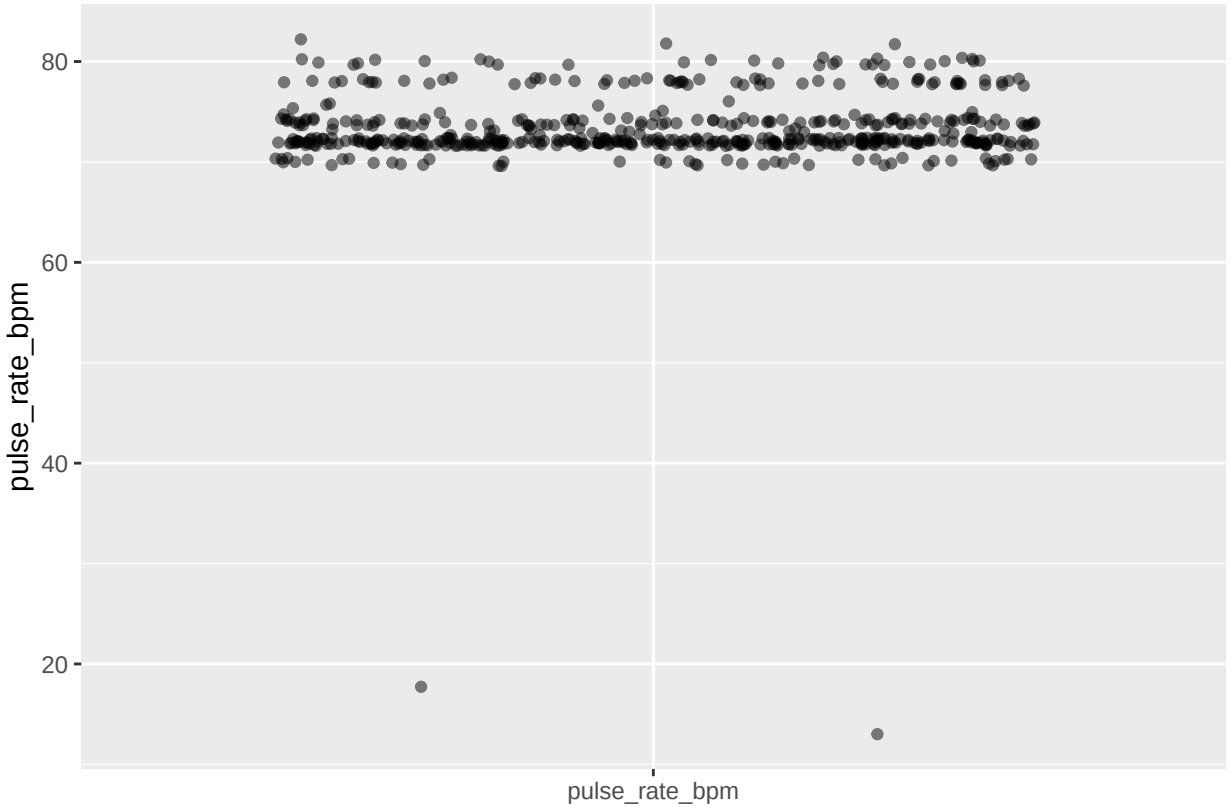
```
data[data$vit_d3_ng_m_l>90,"vit_d3_ng_m_l"] = NA

#Plot the data again

data %>%
  ggplot(aes(x = "vit_d3_ng_m_l", y = vit_d3_ng_m_l)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```

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



Pulse rate It is reported that the normal pulse rate goes from 60 to 100 bpm. Then, I will remove the observations that fall outside of that range

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

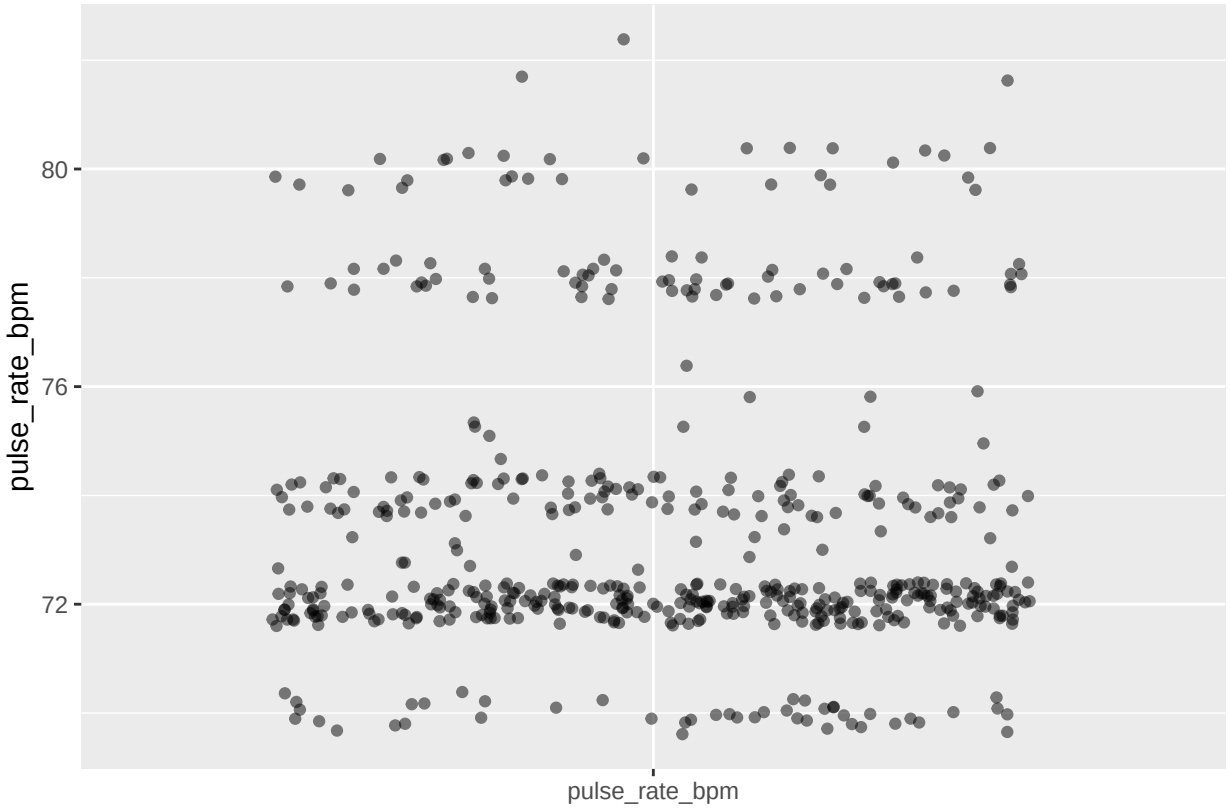



```
data %>%
  filter(pulse_rate_bpm < 60) %>%
  pull(patient_file_no)
```

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

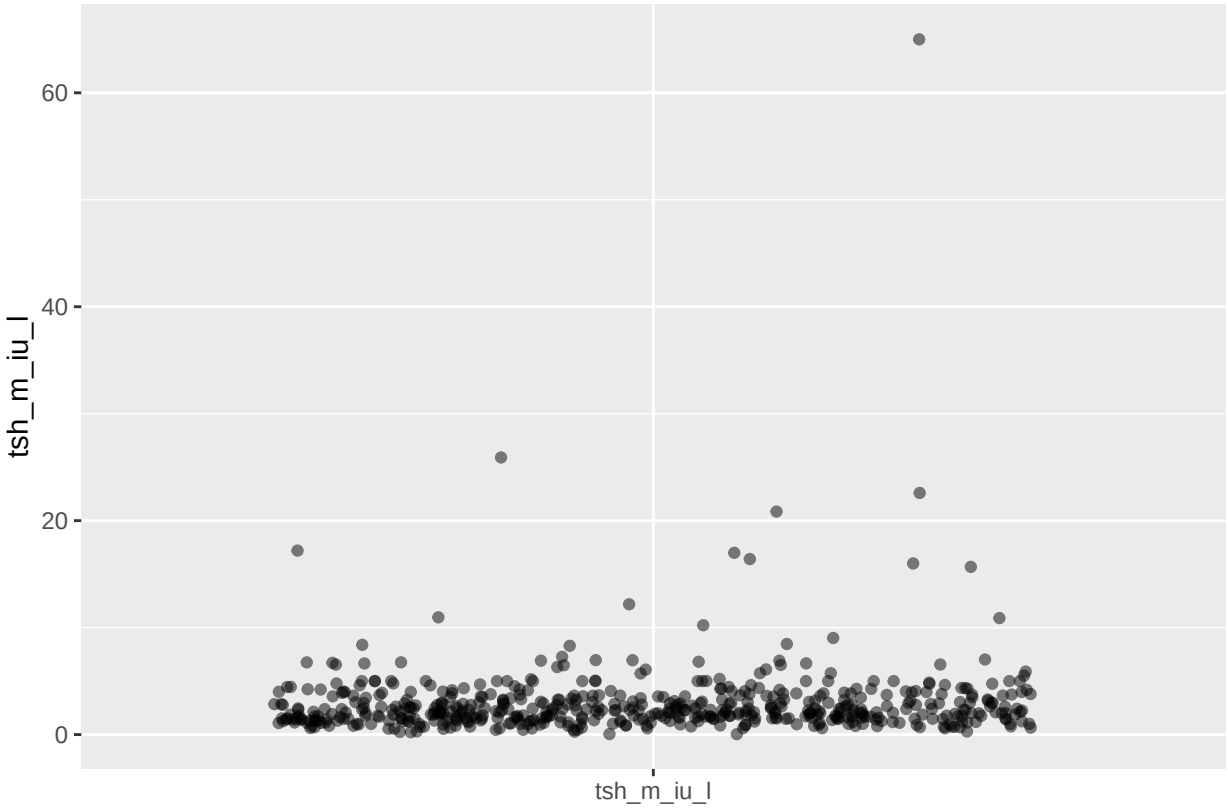
```
data[data$pulse_rate_bpm < 60,"pulse_rate_bpm"] = NA

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

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



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

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



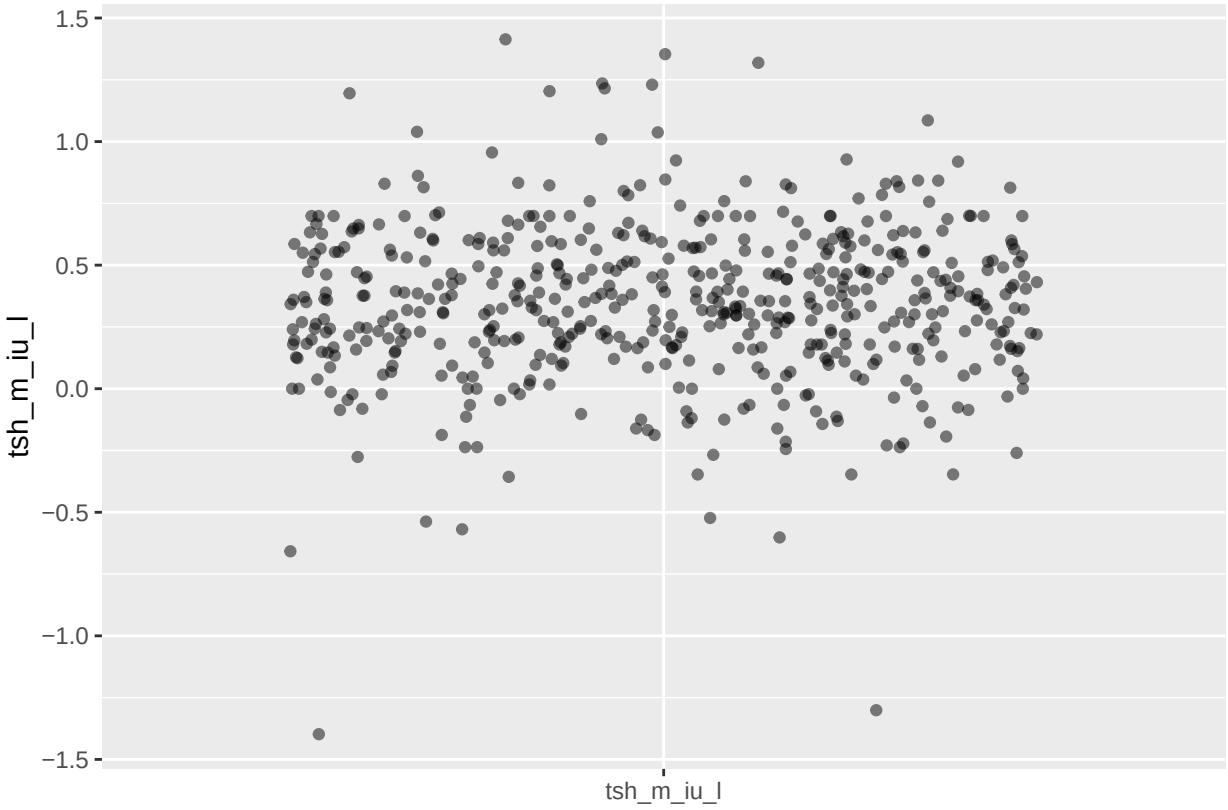
There seems to be an outlier as well, since TSH levels in women with PCOS is around 6.4 ±4.2 mIU/L.

```
data %>%
  filter(tsh_m_iu_l > 40) %>%
  pull(patient_file_no)
```

```
## [1] "38"
```

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

data[data$tsh_m_iu_l  > 40,"tsh_m_iu_l" ] = NA

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

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

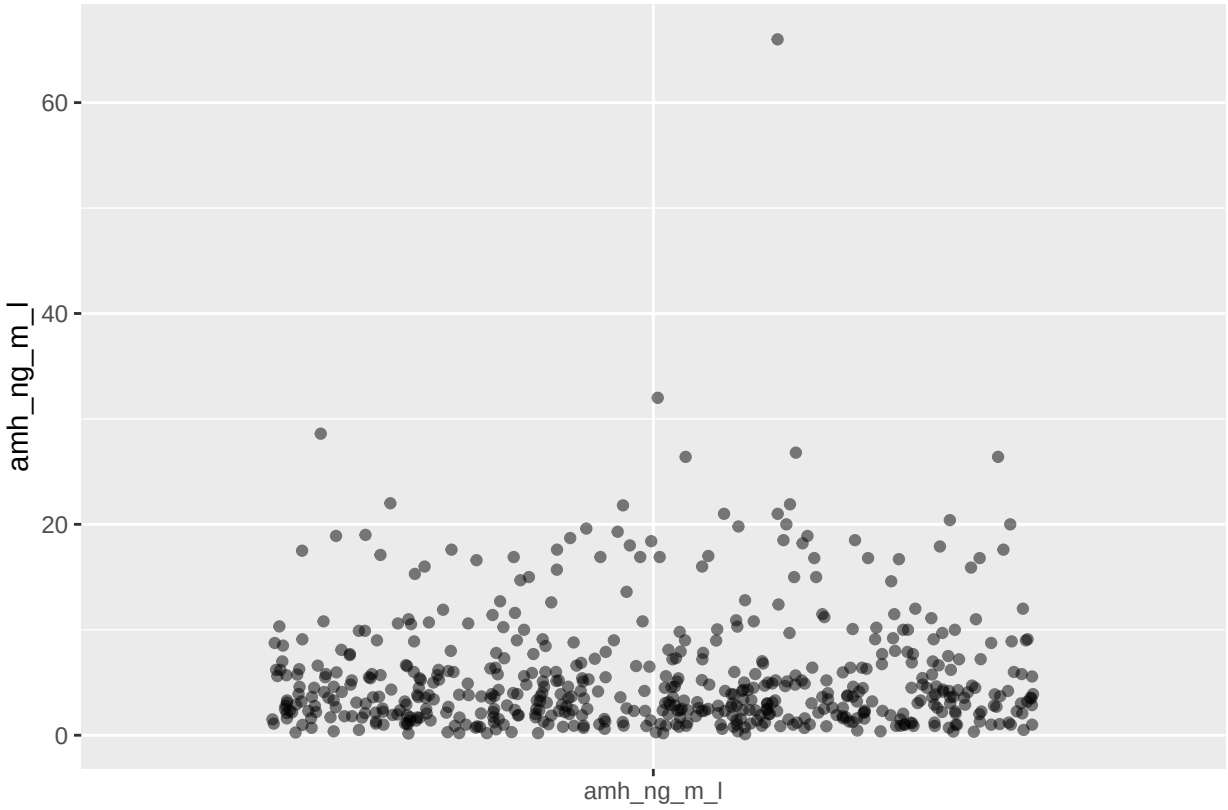
```
## Warning: Removed 1 rows containing missing values (‘geom_point()‘).
```



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

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

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




```
data %>%
  filter(amh_ng_m_l > 40) %>%
  pull(patient_file_no)
```

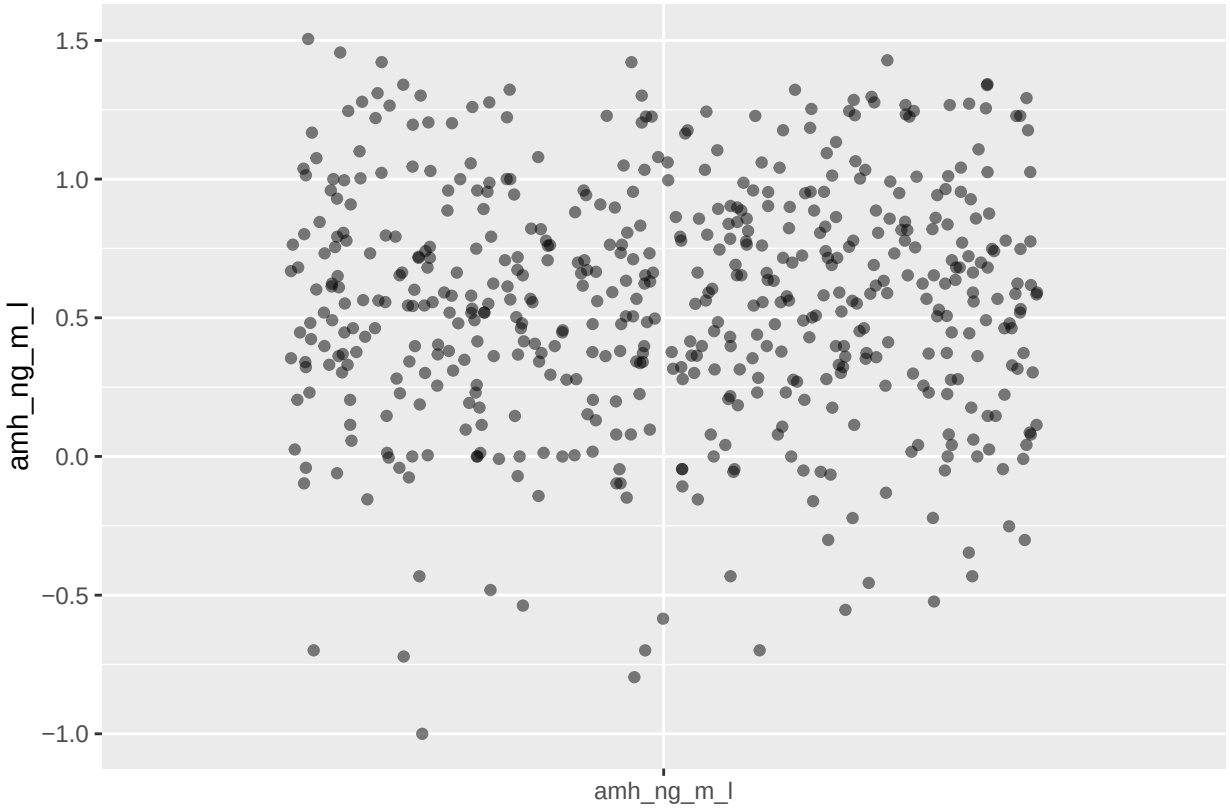
```
## [1] "268"
```

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

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.There are other values that seem to be too high, but I will only remove the outlier that is clearly separated from the rest of the observations.

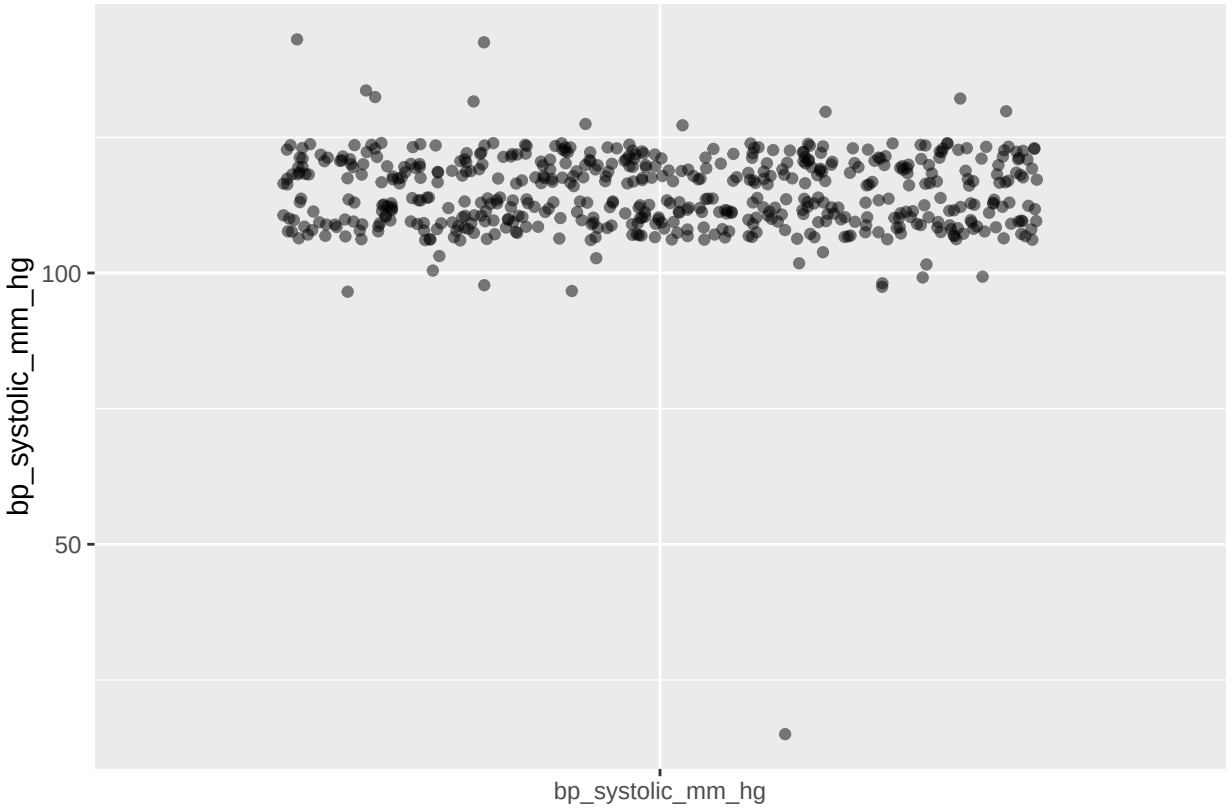
```
data[data$patient_file_no == 268,"amh_ng_m_l" ] = NA

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

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

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

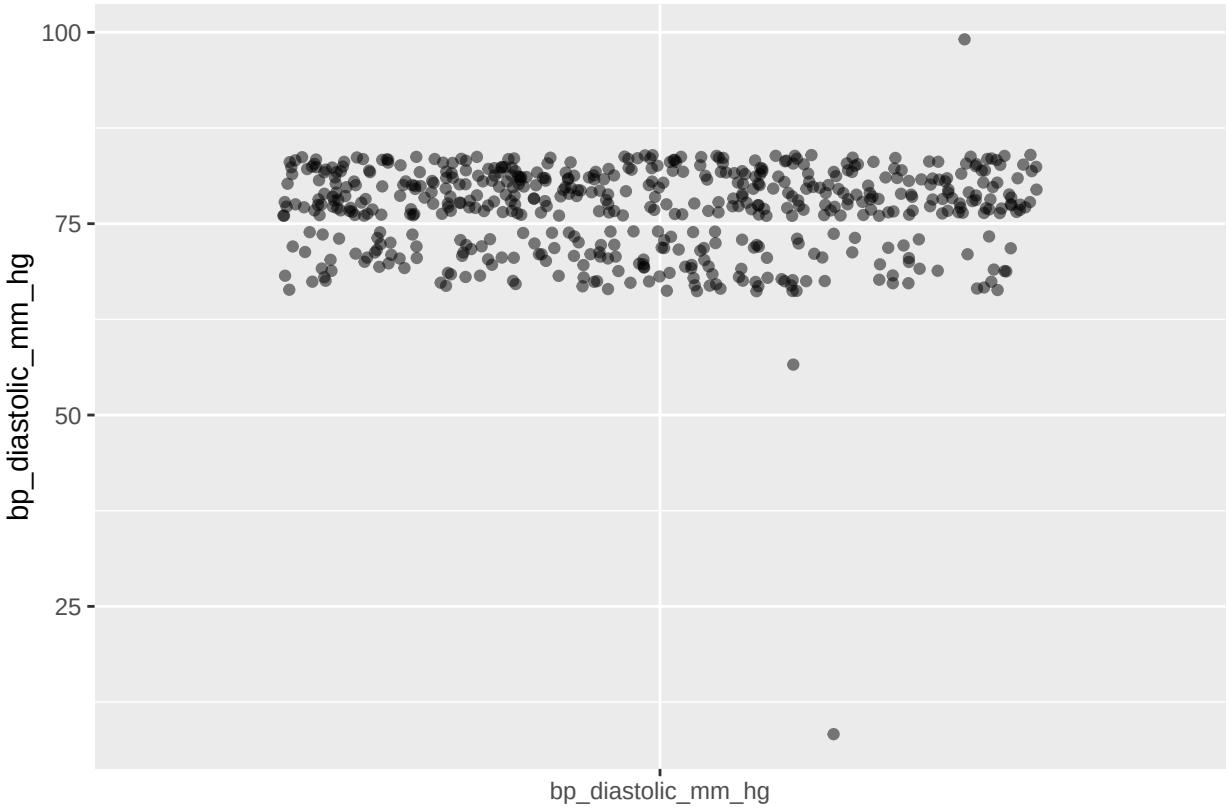


Blood presure Now, I will look for outliers in the blood presure.

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




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



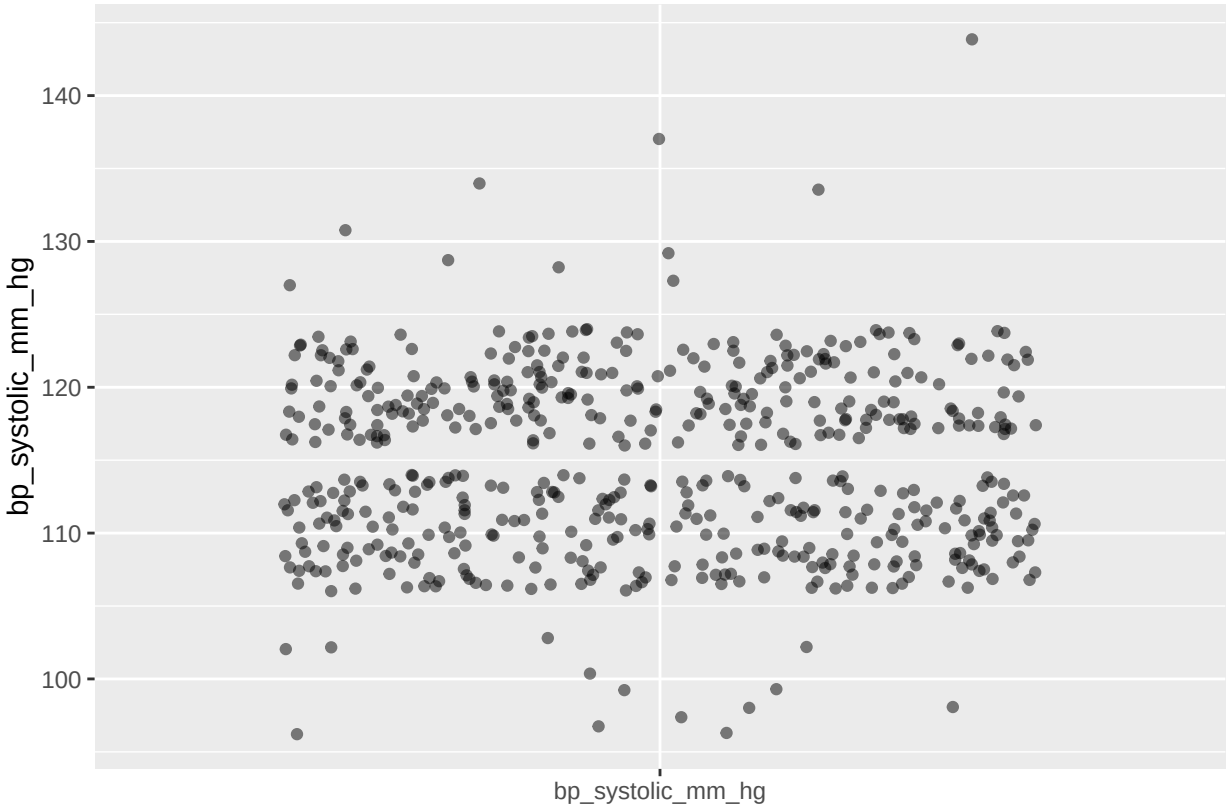
We can observe that there are two different atypical patients with a very odd blood presure. Both of them have a diastolic or systolic blood presure of almost 0 mm/Hg, which is impossible for a living human being. Then, I will set both of them to NA.

```
data %>%
  filter(bp_diastolic_mm_hg < 50 |bp_systolic_mm_hg < 50) %>%
  pull(patient_file_no)
```

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

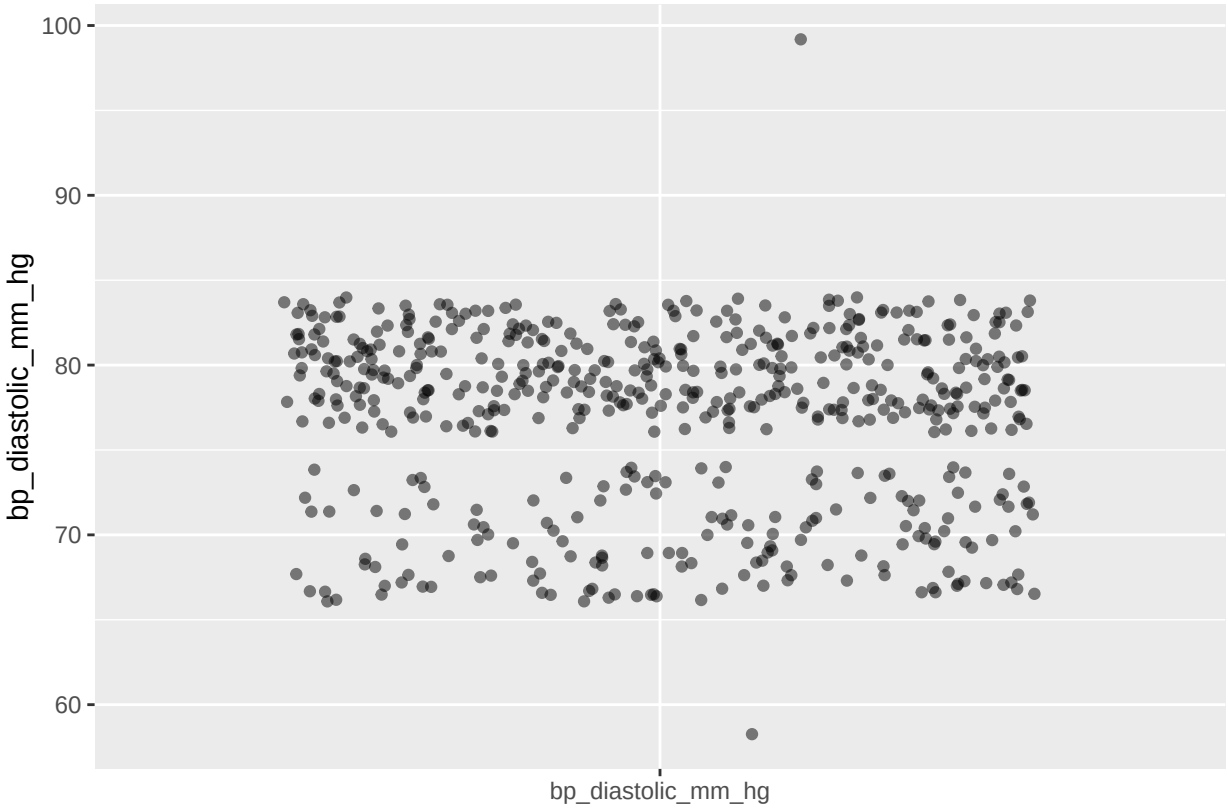
```
#I will flag this patient in case it pops out somewhere else in the analysis.
data[data$bp_diastolic_mm_hg  < 15,"bp_diastolic_mm_hg" ] = NA
data[data$bp_systolic_mm_hg  < 15,"bp_systolic_mm_hg" ] = NA

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

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



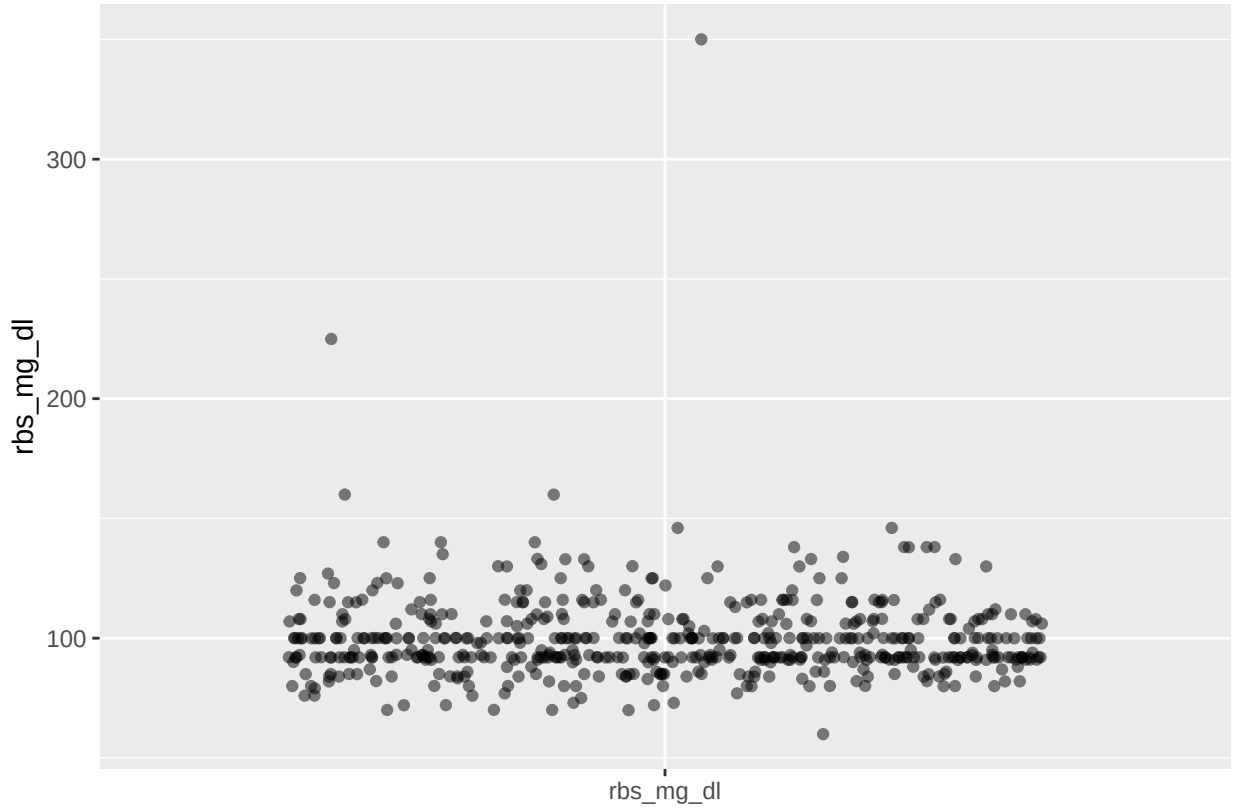

```
data %>%
  ggplot(aes(x = "bp_diastolic_mm_hg", y = bp_diastolic_mm_hg)) +
  geom_jitter(alpha = 0.5) +
  xlab("")
```

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



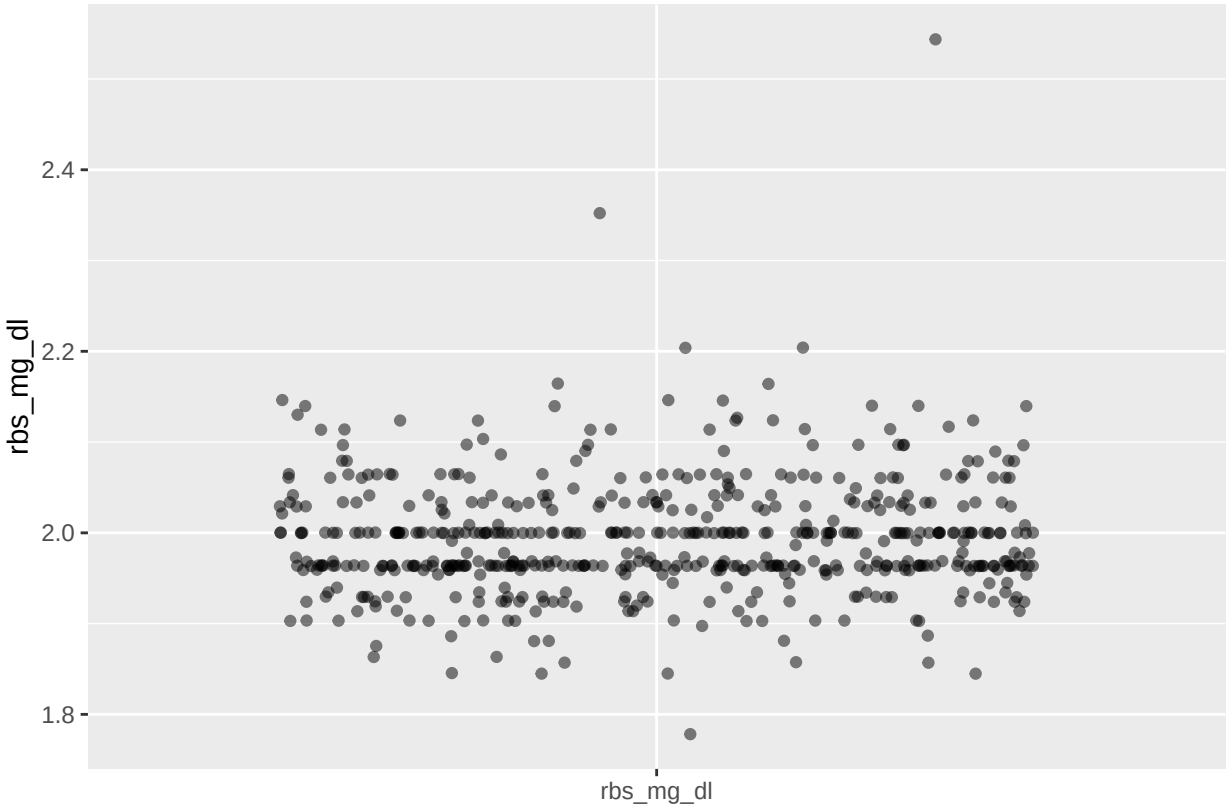
Random blood sugar (glucose) test Now, I will look for outliers in the random blood sugar (glucose) test.

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



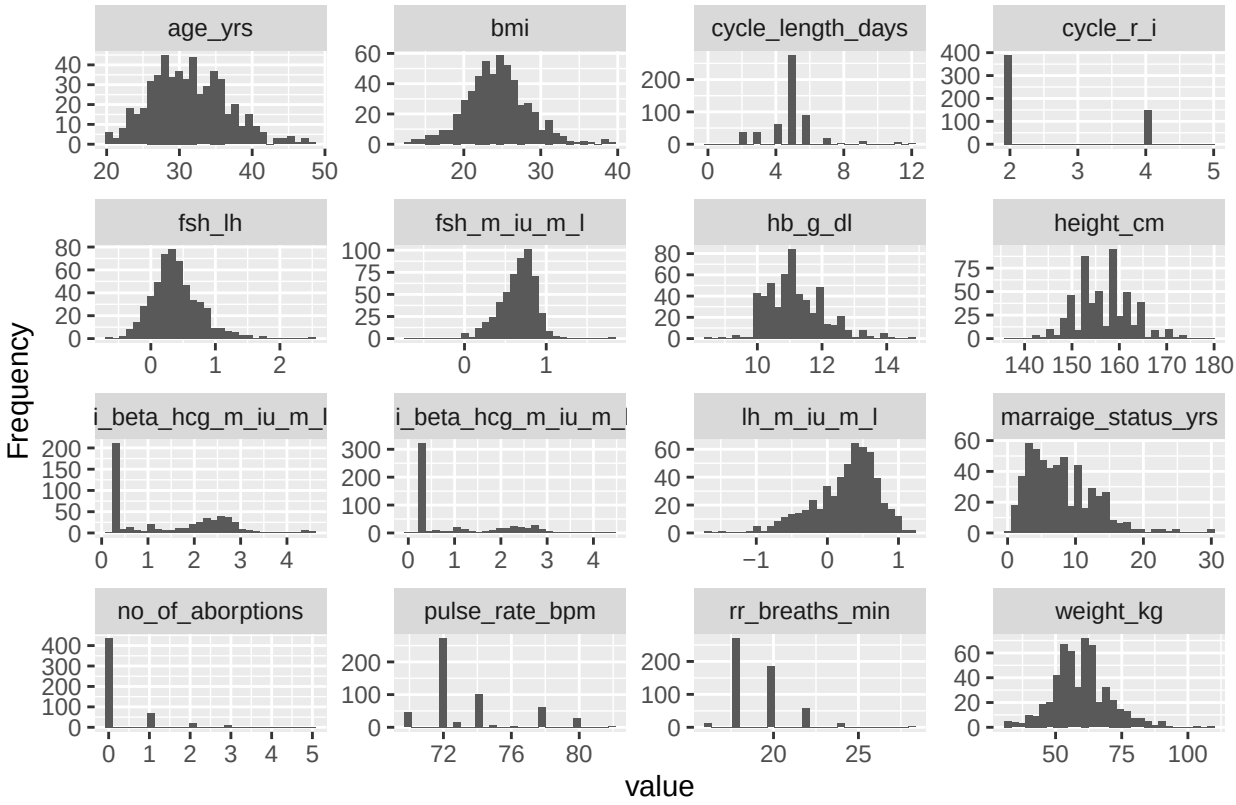
According to the literature, glucose levels can go as up as the ones that are observed. This would likely imply the existence of a syndrome, as well as many physiological consequences. Since this value is then biologically possible, I will keep it. However, I will log-transform the variable.

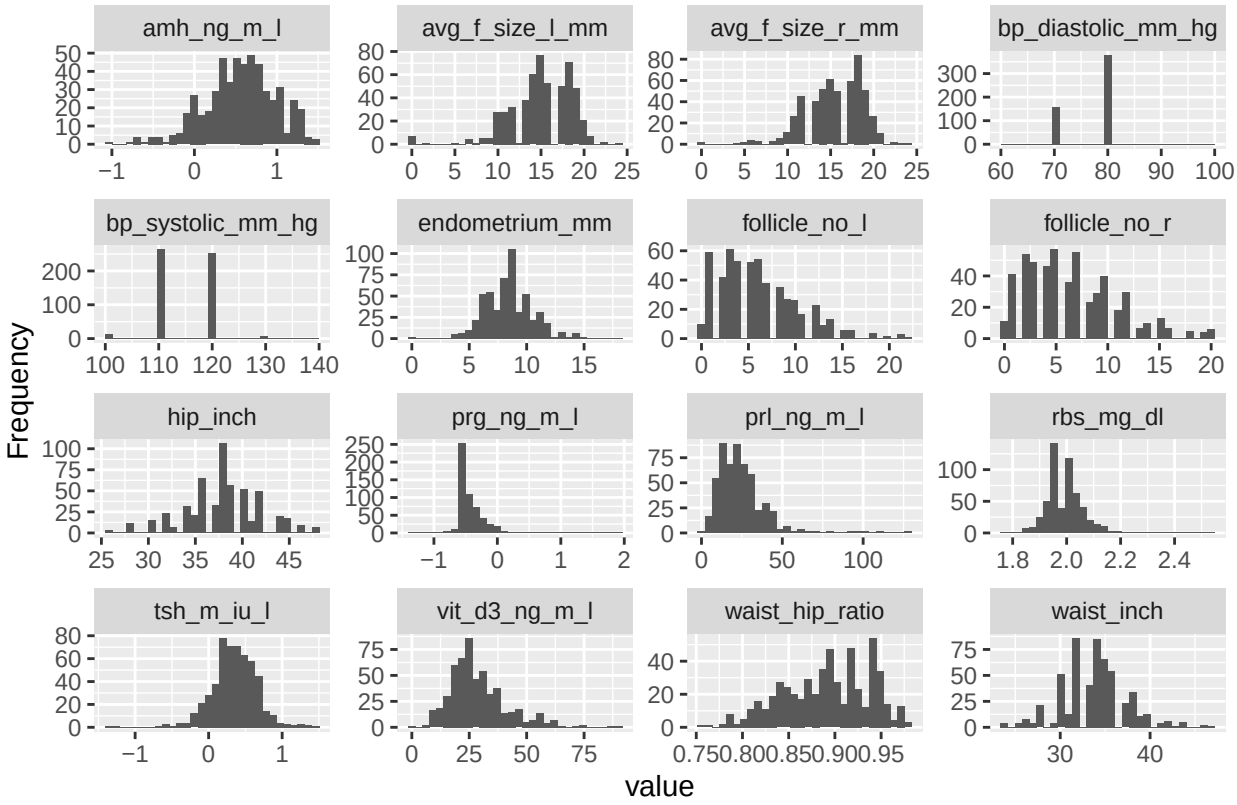
```
#Transform the variable
data = data %>%
  mutate(rbs_mg_dl = log10(rbs_mg_dl))

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



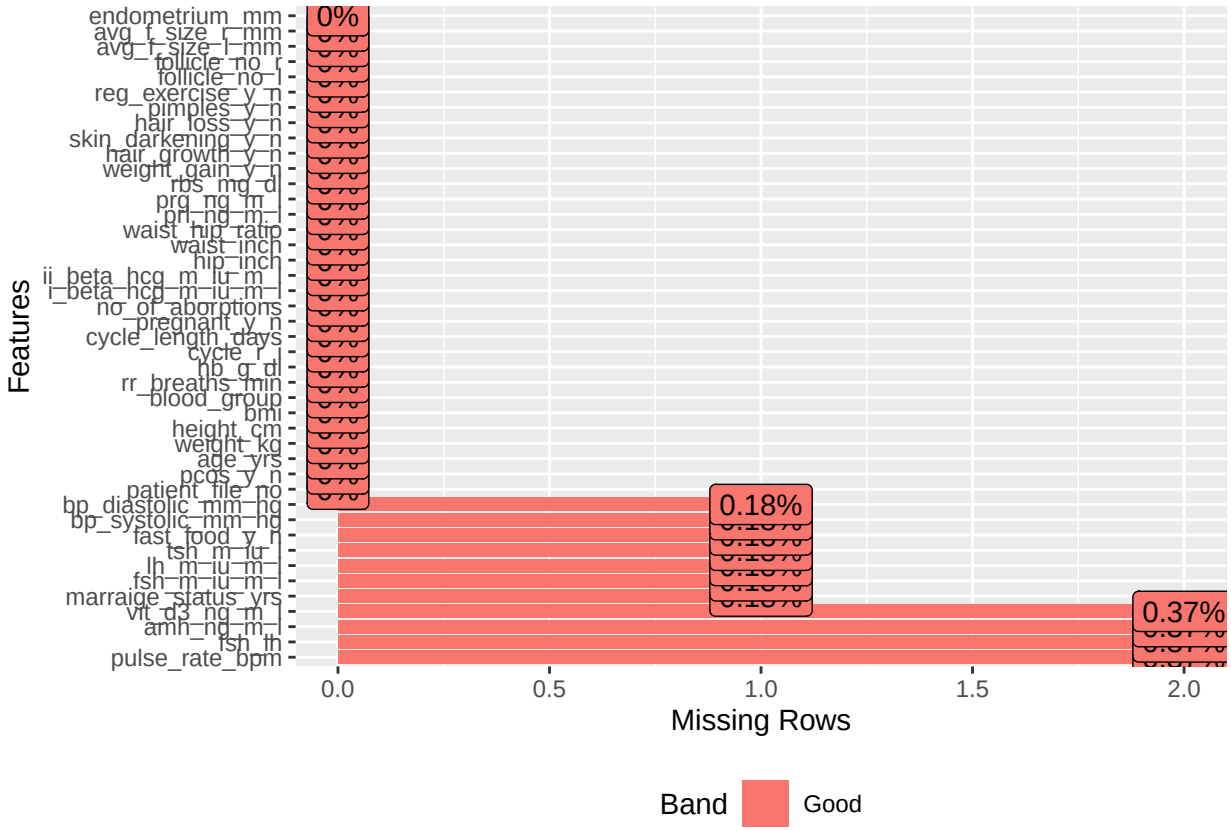
Re plot the distribution of the variables Next, I will re plot the variables to see how the distribution of the variables changed.

```
plot_histogram(data)
```


age_yrs
bmi
cycle_length_days
cycle_r_i
fsh_lh
fsh_m_iu_m_l
hb_g_dl
height_cm
i_beta_hcg_m_iu_m_l
i_beta_hcg_m_iu_m_l
lh_m_iu_m_l
marraige_status_yrs
no_of_aborptions
pulse_rate_bpm
rr_breaths_min
weight_kg
Frequency
value



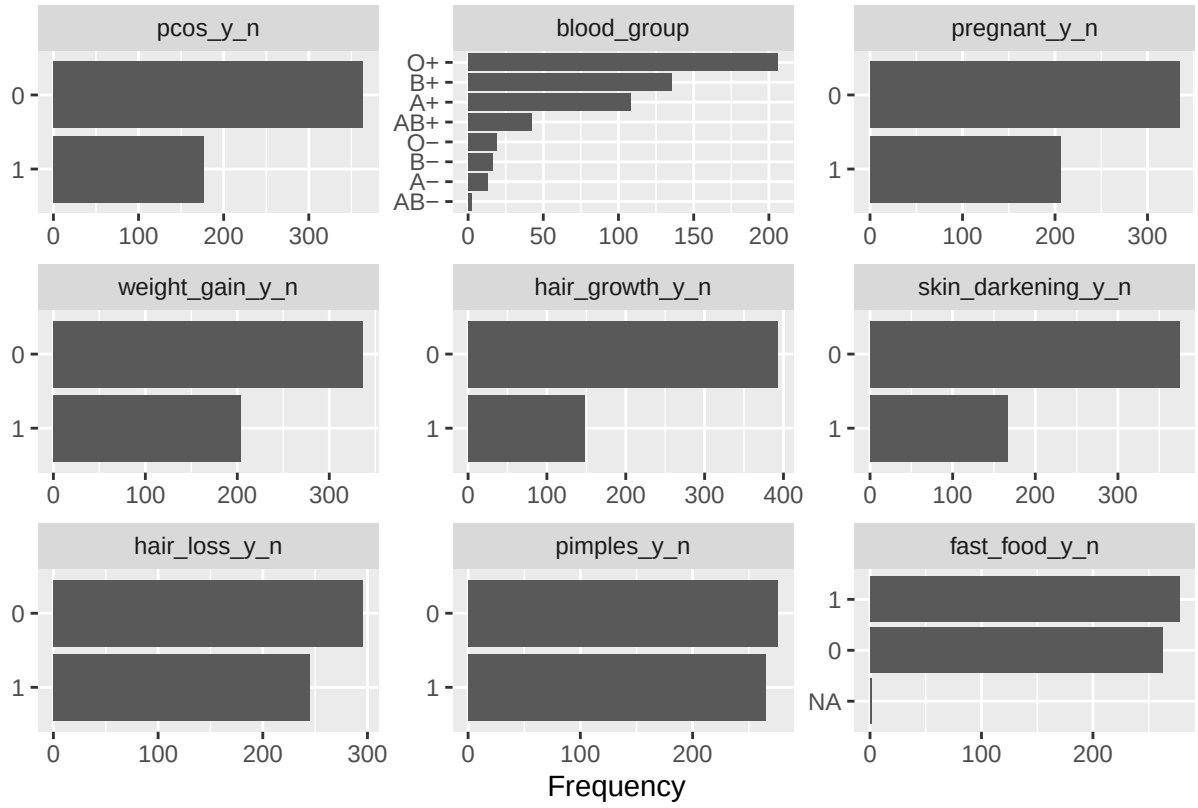

```
plot_missing(data)
```

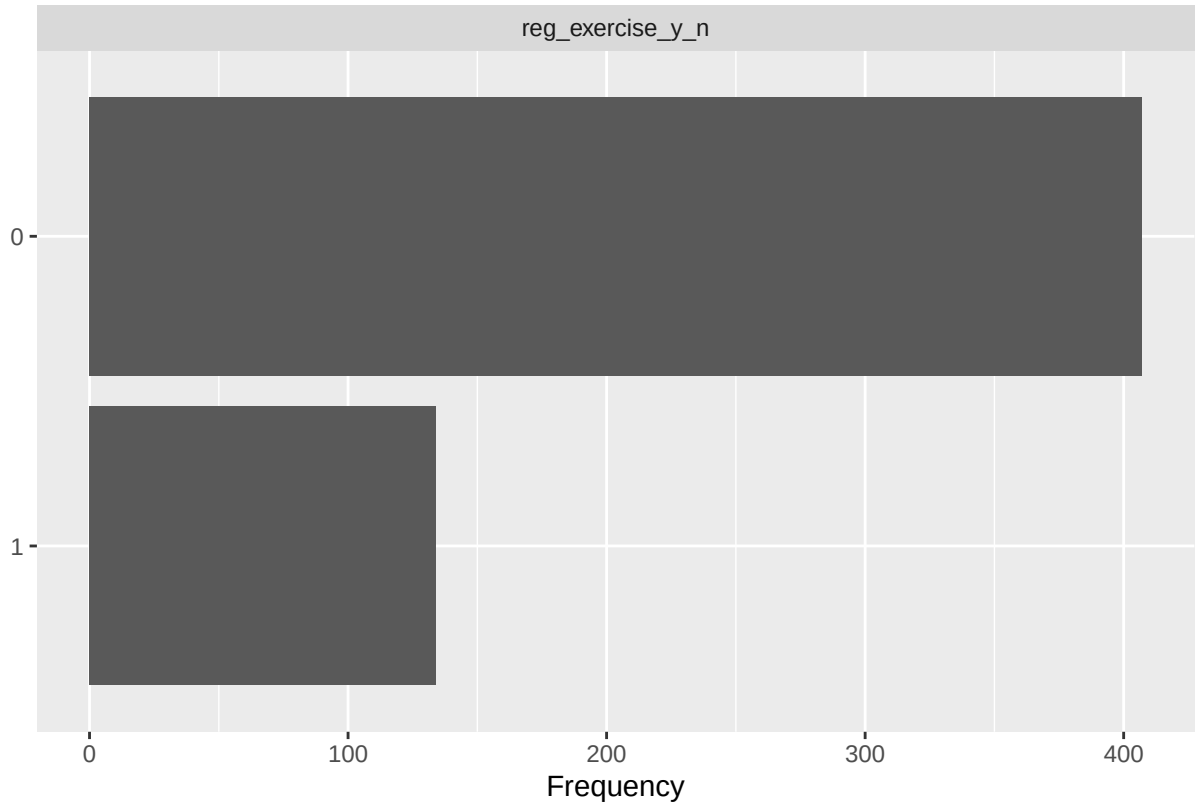


Explore variation of categorical variables

Now, I will explore the variability of the categorical variables.

```
plot_bar(data %>%
           select(-patient_file_no))
```


pcos_y_n
0
1
0 100 200 300
blood_group
O+
B+
A+
AB+
O−
B−
A−
AB−
0 50 100 150 200
pregnant_y_n
0
1
0 100 200 300
weight_gain_y_n
0
1
0 100 200 300
hair_growth_y_n
0
1
0 100 200 300 400
skin_darkening_y_n
0
1
0 100 200 300
hair_loss_y_n
0
1
0 100 200 300
pimples_y_n
0
1
0 100 200
fast_food_y_n
1
0
NA
0 100 200
Frequency

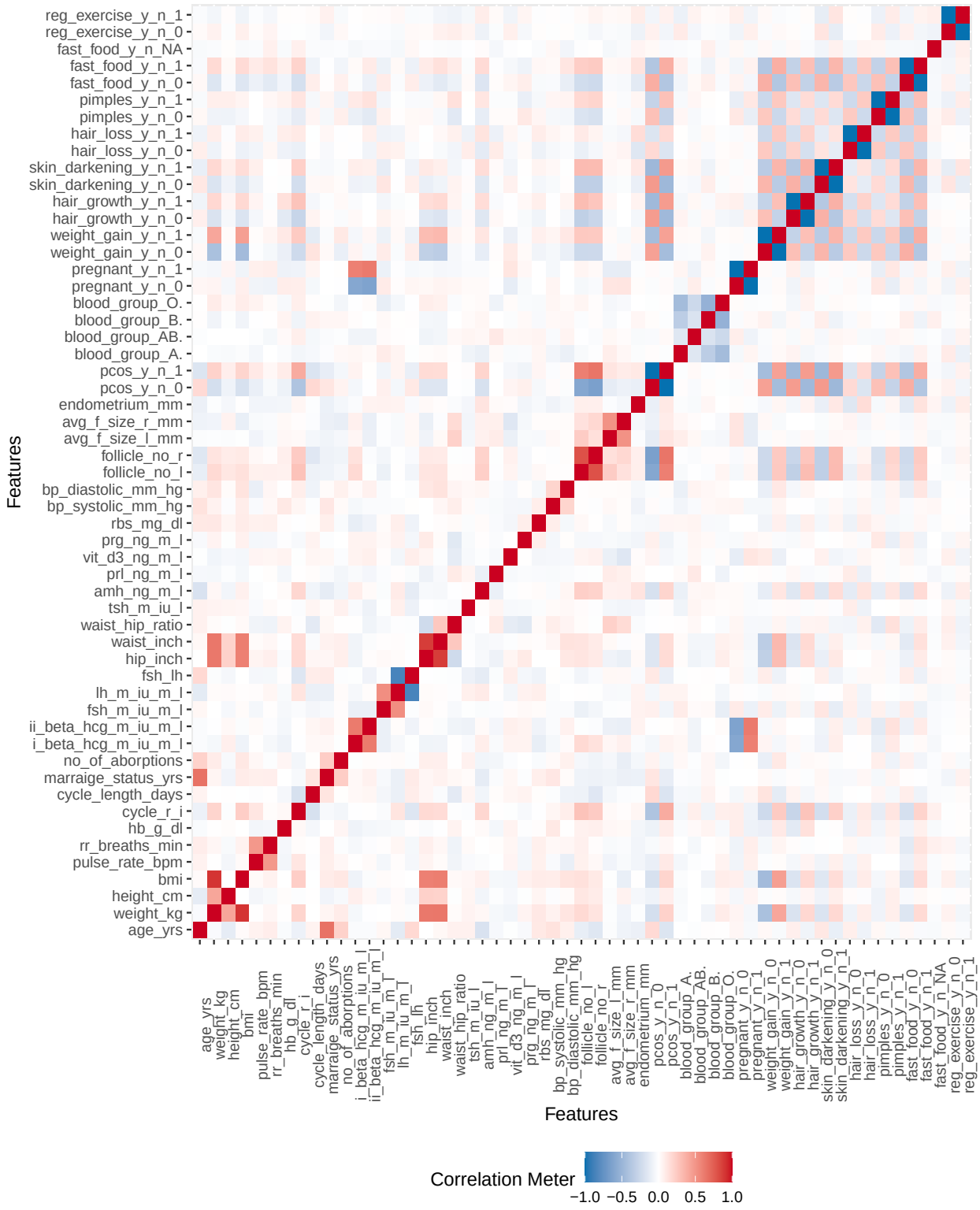


We can observe that none of the variables show an important lack of variation. Then, I won't remove any of them.

Explore covariation

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

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



At this stage, I can observe correlation of some variables that call my attention. I can see some degree of correlation between discrete variables, such as skin darkening, hair growth, weight gain and PCOS. Also, there is correlation between some continuous variables, such as waist and hip, FSH and LH, BMI and weight (which is expected), etc. I will keep this in mind in the future of the analysis, but I won't remove any

variable at this point based on this criteria. I expect these correlations to be addressed in future steps of my project where I implement feature selection.

Save clean object

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

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

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] knitr_1.41         DataExplorer_0.8.2 janitor_2.1.0      readxl_1.4.1
##  [5] here_1.0.1         forcats_0.5.2      stringr_1.5.0      dplyr_1.0.10
##  [9] purrr_0.3.5        readr_2.1.3        tidyr_1.2.1        tibble_3.1.8
## [13] ggplot2_3.4.0      tidyverse_1.3.2
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.9          lubridate_1.9.0     assertthat_0.2.1
##  [4] rprojroot_2.0.3     digest_0.6.31       utf8_1.2.2
##  [7] plyr_1.8.8          R6_2.5.1            cellranger_1.1.0
## [10] backports_1.4.1     reprex_2.0.2        evaluate_0.19
## [13] highr_0.9           httr_1.4.4          pillar_1.8.1
## [16] rlang_1.0.6         googlesheets4_1.0.1 rstudioapi_0.14
## [19] data.table_1.14.6   rmarkdown_2.19      labeling_0.4.2
## [22] googledrive_2.0.0   htmlwidgets_1.6.0   igraph_1.3.5
## [25] munsell_0.5.0       broom_1.0.2         compiler_4.2.2
```

```
## [28] modelr_0.1.10       xfun_0.35           pkgconfig_2.0.3
## [31] htmltools_0.5.4     tidyselect_1.2.0    gridExtra_2.3
## [34] fansi_1.0.3         crayon_1.5.2        tzdb_0.3.0
## [37] dbplyr_2.2.1        withr_2.5.0         grid_4.2.2
## [40] jsonlite_1.8.4      gtable_0.3.1        lifecycle_1.0.3
## [43] DBI_1.1.3           magrittr_2.0.3      scales_1.2.1
## [46] cli_3.4.1           stringi_1.7.8       reshape2_1.4.4
## [49] farver_2.1.1        fs_1.5.2            snakecase_0.11.0
## [52] xml2_1.3.3          ellipsis_0.3.2      generics_0.1.3
## [55] vctrs_0.5.1         tools_4.2.2         glue_1.6.2
## [58] networkD3_0.4       hms_1.1.2           parallel_4.2.2
## [61] fastmap_1.1.0       yaml_2.3.6          timechange_0.1.1
## [64] colorspace_2.0-3    gargle_1.2.1        rvest_1.0.3
## [67] haven_2.5.1
```