Polycystic ovary syndrome diagnosis project

Magdalena Sottanellli

2022-10-18

Introduction

Polycystic ovary syndrome (PCOS) is a common condition that affects how woman's ovaries work during the reproductive years. Women with PCOS usually suffer from irregular periods, hormonal imbalances (high level of "male" hormone androgen) and policystic ovaries. The exact cause of the disease is still unknown, as well as a cure for it. Only the symptoms can be treated. The aim of this project is to create detection and prediction model by using different machine learning algorithms.

About PCOS data set

The data set used in the project is available on KAGGLE and owned by Prasoon Kottarathil. The dataset contains all physical and clinical parameters to determine PCOS and infertility related issues. Data is collected from 10 different hospitals across Kerala in India. Original data set is saved in two different files - infertility and without-infertility patients. There are 51 columns in total, of which 41 are different parameters that are used to describee our target column - PCOS (Y/N).

Important notes regarding data set: - The unit used is feet to cm - Blood pressure entered as systolic and diastolic separately - RBS stands for Random glucose test - Beta-HCG cases are mentioned as Case I and II - Blood Group indications: A+ = 11, A- = 12, B+ = 13, B- = 14, O+ =15, O- = 16, AB+ =17, AB- = 18

Loading and cleaning the data

First we load libraries needed to perform analysis as follow:

```r
options(digits = 3)
library(tidyverse)
library(caret)
library(data.table)
library(ggplot2)
library(lubridate)
library(dslabs)
library(readxl)
library(readr)
library(corrplot)
```

Data has been uploaded and merged by the column 'Sl. No' with below code:

```r
path_1 <- "C:/Users/A529462/Desktop/R files/PCOS project/PCOS_data_without_infertility.xlsx"
pcos_1 <- read_excel(path_1, sheet = "Full_new")
```

```
## New names:
## * ‘‘ -> ‘...45‘
```

```r
path_2 <- "C:/Users/A529462/Desktop/R files/PCOS project/PCOS_infertility.csv"
pcos_2 <- read_csv(path_2)
```

```
## Rows: 541 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): AMH(ng/mL)
## dbl (5): Sl. No, Patient File No., PCOS (Y/N), I   beta-HCG(mIU/mL), II    b...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
pcos <- merge(pcos_1, pcos_2, by = 'Sl. No')
```

Next, we cleaned data from duplicate columns.

```r
pcos <- subset (pcos, select =  -c(`Patient File No..y`,`I   beta-HCG(mIU/mL).y`,
                                   `PCOS (Y/N).y`, `I   beta-HCG(mIU/mL).y`,
                                   `II    beta-HCG(mIU/mL).y`, `AMH(ng/mL).y`, `...45`))
```

After checking the data types of the columns, some columns needed to be changed to numeric.

```r
str(pcos)
```

```
## 'data.frame':    541 obs. of  44 variables:
##  $ Sl. No                  : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ Patient File No..x      : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ PCOS (Y/N).x            : num  0 0 1 0 0 0 0 0 0 0 ...
##  $ Age (yrs)               : num  28 36 33 37 25 36 34 33 32 36 ...
##  $ Weight (Kg)             : num  44.6 65 68.8 65 52 74.1 64 58.5 40 52 ...
##  $ Height(Cm)              : num  152 162 165 148 161 ...
##  $ BMI                     : num  19.3 24.9 25.3 29.7 20.1 ...
##  $ Blood Group             : num  15 15 11 13 11 15 11 13 11 15 ...
##  $ Pulse rate(bpm)         : num  78 74 72 72 72 78 72 72 72 80 ...
##  $ RR (breaths/min)        : num  22 20 18 20 18 28 18 20 18 20 ...
##  $ Hb(g/dl)                : num  10.5 11.7 11.8 12 10 ...
##  $ Cycle(R/I)              : num  2 2 2 2 2 2 2 2 2 4 ...
##  $ Cycle length(days)      : num  5 5 5 5 5 5 5 5 5 2 ...
##  $ Marraige Status (Yrs)   : num  7 11 10 4 1 8 2 13 8 4 ...
##  $ Pregnant(Y/N)           : num  0 1 1 0 1 1 0 1 0 0 ...
##  $ No. of aborptions       : num  0 0 0 0 0 0 0 2 1 0 ...
##  $ I   beta-HCG(mIU/mL).x  : num  1.99 60.8 494.08 1.99 801.45 ...
##  $ II    beta-HCG(mIU/mL).x: chr  "1.99" "1.99" "494.08" "1.99" ...
##  $ FSH(mIU/mL)             : num  7.95 6.73 5.54 8.06 3.98 3.24 2.85 4.86 3.76 2.8 ...
##  $ LH(mIU/mL)              : num  3.68 1.09 0.88 2.36 0.9 1.07 0.31 3.07 3.02 1.51 ...
##  $ FSH/LH                  : num  2.16 6.17 6.3 3.42 4.42 ...
##  $ Hip(inch)               : num  36 38 40 42 37 44 39 44 39 40 ...
##  $ Waist(inch)             : num  30 32 36 36 30 38 33 38 35 38 ...
##  $ Waist:Hip Ratio         : num  0.833 0.842 0.9 0.857 0.811 ...
##  $ TSH (mIU/L)             : num  0.68 3.16 2.54 16.41 3.57 ...
##  $ AMH(ng/mL).x            : chr  "2.07" "1.53" "6.63" "1.22" ...
##  $ PRL(ng/mL)              : num  45.2 20.1 10.5 36.9 30.1 ...
##  $ Vit D3 (ng/mL)          : num  17.1 61.3 49.7 33.4 43.8 52.4 42.7 38 21.8 27.7 ...
##  $ PRG(ng/mL)              : num  0.57 0.97 0.36 0.36 0.38 0.3 0.46 0.26 0.3 0.25 ...
##  $ RBS(mg/dl)              : num  92 92 84 76 84 76 93 91 116 125 ...
##  $ Weight gain(Y/N)        : num  0 0 0 0 0 1 0 1 0 0 ...
##  $ hair growth(Y/N)        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Skin darkening (Y/N)    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Hair loss(Y/N)          : num  0 0 1 0 1 1 0 0 0 0 ...
##  $ Pimples(Y/N)            : num  0 0 1 0 0 0 0 0 0 0 ...
##  $ Fast food (Y/N)         : num  1 0 1 0 0 0 0 0 0 0 ...
##  $ Reg.Exercise(Y/N)       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BP _Systolic (mmHg)     : num  110 120 120 120 120 110 120 120 120 110 ...
##  $ BP _Diastolic (mmHg)    : num  80 70 80 70 80 70 80 80 80 80 ...
##  $ Follicle No. (L)        : num  3 3 13 2 3 9 6 7 5 1 ...
##  $ Follicle No. (R)        : num  3 5 15 2 4 6 6 6 7 1 ...
##  $ Avg. F size (L) (mm)    : num  18 15 18 15 16 16 15 15 17 14 ...
##  $ Avg. F size (R) (mm)    : num  18 14 20 14 14 20 16 18 17 17 ...
##  $ Endometrium (mm)        : num  8.5 3.7 10 7.5 7 8 6.8 7.1 4.2 2.5 ...
```

```
pcos$`II    beta-HCG(mIU/mL).x`<- as.numeric(pcos$`II    beta-HCG(mIU/mL).x`)
pcos$`AMH(ng/mL).x` <- as.numeric(pcos$`AMH(ng/mL).x`)
```

Some of the columns (Fast food (Y/N),Marraige Status (Yrs), II beta-HCG(mIU/mL).x, AMH(ng/mL)) have NA values. They have been replaced with the median value of the column.

```
sapply(pcos, function(x) sum(is.na(x)))
```

```
##                   Sl. No       Patient File No..x             PCOS (Y/N).x 
##                        0                        0                        0 
##                Age (yrs)              Weight (Kg)               Height(Cm) 
##                        0                        0                        0 
##                      BMI              Blood Group          Pulse rate(bpm) 
##                        0                        0                        0 
##         RR (breaths/min)                 Hb(g/dl)               Cycle(R/I) 
##                        0                        0                        0 
##       Cycle length(days)    Marraige Status (Yrs)            Pregnant(Y/N) 
##                        0                        1                        0 
##        No. of aborptions   I   beta-HCG(mIU/mL).x II    beta-HCG(mIU/mL).x 
##                        0                        0                        1 
##              FSH(mIU/mL)               LH(mIU/mL)                   FSH/LH 
##                        0                        0                        0 
##                Hip(inch)              Waist(inch)          Waist:Hip Ratio 
##                        0                        0                        0 
##              TSH (mIU/L)             AMH(ng/mL).x               PRL(ng/mL) 
##                        0                        1                        0 
##           Vit D3 (ng/mL)               PRG(ng/mL)               RBS(mg/dl) 
##                        0                        0                        0
##         Weight gain(Y/N)         hair growth(Y/N)     Skin darkening (Y/N)
##                        0                        0                        0
##           Hair loss(Y/N)             Pimples(Y/N)          Fast food (Y/N)
##                        0                        0                        1
##        Reg.Exercise(Y/N)      BP _Systolic (mmHg)     BP _Diastolic (mmHg)
##                        0                        0                        0
##         Follicle No. (L)         Follicle No. (R)     Avg. F size (L) (mm)
##                        0                        0                        0
##     Avg. F size (R) (mm)         Endometrium (mm)
##                        0                        0
```

```
pcos$`Fast food (Y/N)`[is.na(pcos$`Fast food (Y/N)`)] <- median(pcos$`Fast food (Y/N)`, na.rm = T)
pcos$`Marraige Status (Yrs)`[is.na(pcos$`Marraige Status (Yrs)`)] <- median(pcos$`Marraige Status (Yrs)
pcos$`II    beta-HCG(mIU/mL).x`[is.na(pcos$`II    beta-HCG(mIU/mL).x`)] <- median(pcos$`II    beta-HCG(m
pcos$`AMH(ng/mL).x`[is.na(pcos$`AMH(ng/mL).x`)] <- median(pcos$`AMH(ng/mL).x`, na.rm = T)
```

We also removed two columns with little information (Sl. No, Patient File No.), hence they only contain information about file number of a patient.

Exploratory data analysis

Lets have a look at the descriptive statistics, by using below code:

```
summary(pcos)
```

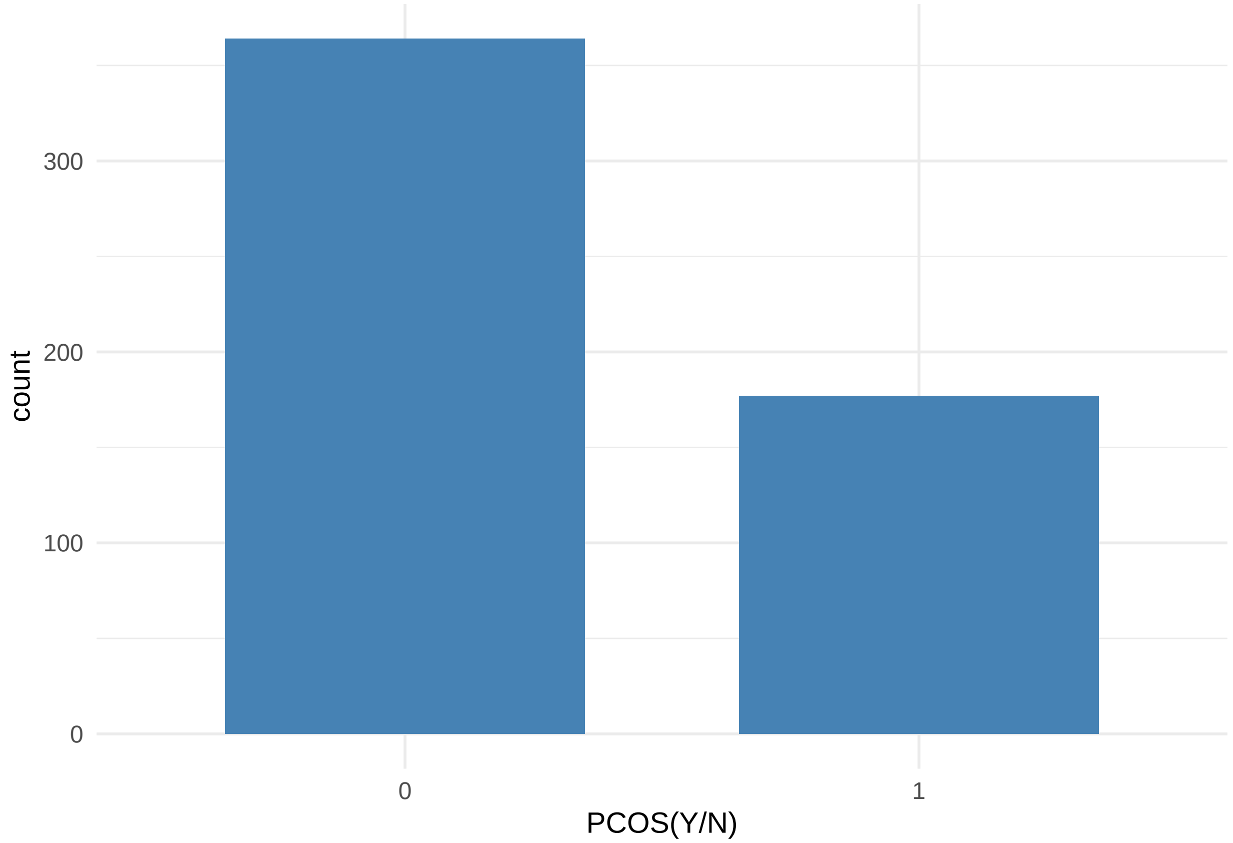
```
##      Sl. No    Patient File No..x  PCOS (Y/N).x     Age (yrs)
##  Min.   :  1   Min.   :  1        Min.   :0.000   Min.   :20.0
##  1st Qu.:136   1st Qu.:136        1st Qu.:0.000   1st Qu.:28.0
##  Median :271   Median :271        Median :0.000   Median :31.0
##  Mean   :271   Mean   :271        Mean   :0.327   Mean   :31.4
##  3rd Qu.:406   3rd Qu.:406        3rd Qu.:1.000   3rd Qu.:35.0
##  Max.   :541   Max.   :541        Max.   :1.000   Max.   :48.0
##   Weight (Kg)      Height(Cm)       BMI        Blood Group   Pulse rate(bpm)
##  Min.   : 31.0   Min.   :137   Min.   :12.4   Min.   :11.0   Min.   :13.0
##  1st Qu.: 52.0   1st Qu.:152   1st Qu.:21.6   1st Qu.:13.0   1st Qu.:72.0
##  Median : 59.0   Median :156   Median :24.2   Median :14.0   Median :72.0
##  Mean   : 59.6   Mean   :156   Mean   :24.3   Mean   :13.8   Mean   :73.2
##  3rd Qu.: 65.0   3rd Qu.:160   3rd Qu.:26.6   3rd Qu.:15.0   3rd Qu.:74.0
##  Max.   :108.0   Max.   :180   Max.   :38.9   Max.   :18.0   Max.   :82.0
##  RR (breaths/min)    Hb(g/dl)      Cycle(R/I)   Cycle length(days)
##  Min.   :16.0     Min.   : 8.5   Min.   :2.00   Min.   : 0.00
##  1st Qu.:18.0     1st Qu.:10.5   1st Qu.:2.00   1st Qu.: 4.00
##  Median :18.0     Median :11.0   Median :2.00   Median : 5.00
##  Mean   :19.2     Mean   :11.2   Mean   :2.56   Mean   : 4.94
##  3rd Qu.:20.0     3rd Qu.:11.7   3rd Qu.:4.00   3rd Qu.: 5.00
##  Max.   :28.0     Max.   :14.8   Max.   :5.00   Max.   :12.00
##  Marraige Status (Yrs) Pregnant(Y/N)   No. of aborptions I   beta-HCG(mIU/mL).x
##  Min.   : 0.00         Min.   :0.000   Min.   :0.00      Min.   :    1
##  1st Qu.: 4.00         1st Qu.:0.000   1st Qu.:0.00      1st Qu.:    2
##  Median : 7.00         Median :0.000   Median :0.00      Median :   20
##  Mean   : 7.68         Mean   :0.381   Mean   :0.29      Mean   :  665
##  3rd Qu.:10.00         3rd Qu.:1.000   3rd Qu.:0.00      3rd Qu.:  297
##  Max.   :30.00         Max.   :1.000   Max.   :5.00      Max.   :32461
##  II    beta-HCG(mIU/mL).x  FSH(mIU/mL)     LH(mIU/mL)       FSH/LH
##  Min.   :    1            Min.   :   0   Min.   :   0   Min.   :   0
##  1st Qu.:    2            1st Qu.:   3   1st Qu.:   1   1st Qu.:   1
##  Median :    2            Median :   5   Median :   2   Median :   2
##  Mean   :  238            Mean   :  15   Mean   :   6   Mean   :   7
##  3rd Qu.:   98            3rd Qu.:   6   3rd Qu.:   4   3rd Qu.:   4
##  Max.   :25000            Max.   :5052   Max.   :2018   Max.   :1373
##    Hip(inch)   Waist(inch)   Waist:Hip Ratio  TSH (mIU/L)    AMH(ng/mL).x
##  Min.   :26   Min.   :24.0   Min.   :0.756   Min.   : 0.0   Min.   : 0.1
##  1st Qu.:36   1st Qu.:32.0   1st Qu.:0.857   1st Qu.: 1.5   1st Qu.: 2.0
##  Median :38   Median :34.0   Median :0.895   Median : 2.3   Median : 3.7
##  Mean   :38   Mean   :33.8   Mean   :0.892   Mean   : 3.0   Mean   : 5.6
##  3rd Qu.:40   3rd Qu.:36.0   3rd Qu.:0.929   3rd Qu.: 3.6   3rd Qu.: 6.9
##  Max.   :48   Max.   :47.0   Max.   :0.979   Max.   :65.0   Max.   :66.0
##    PRL(ng/mL)    Vit D3 (ng/mL)   PRG(ng/mL)     RBS(mg/dl)  Weight gain(Y/N)
##  Min.   :  0.4   Min.   :   0   Min.   : 0.0   Min.   : 60   Min.   :0.000
##  1st Qu.: 14.5   1st Qu.:  21   1st Qu.: 0.2   1st Qu.: 92   1st Qu.:0.000
##  Median : 21.9   Median :  26   Median : 0.3   Median :100   Median :0.000
##  Mean   : 24.3   Mean   :  50   Mean   : 0.6   Mean   :100   Mean   :0.377
##  3rd Qu.: 29.9   3rd Qu.:  34   3rd Qu.: 0.4   3rd Qu.:107   3rd Qu.:1.000
##  Max.   :128.2   Max.   :6015   Max.   :85.0   Max.   :350   Max.   :1.000
##  hair growth(Y/N) Skin darkening (Y/N) Hair loss(Y/N)   Pimples(Y/N)
##  Min.   :0.000    Min.   :0.000        Min.   :0.000   Min.   :0.00
##  1st Qu.:0.000    1st Qu.:0.000        1st Qu.:0.000   1st Qu.:0.00
##  Median :0.000    Median :0.000        Median :0.000   Median :0.00
##  Mean   :0.274    Mean   :0.307        Mean   :0.453   Mean   :0.49
##  3rd Qu.:1.000    3rd Qu.:1.000        3rd Qu.:1.000   3rd Qu.:1.00
##  Max.   :1.000    Max.   :1.000        Max.   :1.000   Max.   :1.00
##  Fast food (Y/N) Reg.Exercise(Y/N) BP _Systolic (mmHg) BP _Diastolic (mmHg)
##  Min.   :0.000   Min.   :0.000     Min.   : 12         Min.   :  8.0
##  1st Qu.:0.000   1st Qu.:0.000     1st Qu.:110         1st Qu.: 70.0
##  Median :1.000   Median :0.000     Median :110         Median : 80.0
##  Mean   :0.516   Mean   :0.248     Mean   :115         Mean   : 76.9
##  3rd Qu.:1.000   3rd Qu.:0.000     3rd Qu.:120         3rd Qu.: 80.0
##  Max.   :1.000   Max.   :1.000     Max.   :140         Max.   :100.0
##  Follicle No. (L) Follicle No. (R) Avg. F size (L) (mm) Avg. F size (R) (mm)
##  Min.   : 0.00    Min.   : 0.00    Min.   : 0           Min.   : 0.0
##  1st Qu.: 3.00    1st Qu.: 3.00    1st Qu.:13           1st Qu.:13.0
##  Median : 5.00    Median : 6.00    Median :15           Median :16.0
##  Mean   : 6.13    Mean   : 6.64    Mean   :15           Mean   :15.4
##  3rd Qu.: 9.00    3rd Qu.:10.00    3rd Qu.:18           3rd Qu.:18.0
##  Max.   :22.00    Max.   :20.00    Max.   :24           Max.   :24.0
##  Endometrium (mm)
##  Min.   : 0.00
##  1st Qu.: 7.00
##  Median : 8.50
##  Mean   : 8.48
##  3rd Qu.: 9.80
##  Max.   :18.00
```

Next we will look into variables in more details. In general, we can see that in our data set there are two types of variables - categorical and numeric. The answer to categorical variables is yes/no, which in data set is present as 1 and 0 respectively.

To this category belong below columns: PCOS(Y/N), Pregnant(Y/N), Weight_gain(Y/N), hair_growth(Y/N), Skin_darkening(Y/N), Hair_loss(Y/N),Pimples(Y/N), Fast_food(Y/N), Reg_Exercise(Y/N), Blood Group.

Lets start with our target column PCOS(Y/N), which indicates wheater patient has or has not policystic ovary syndrome.

```
ggplot(pcos, aes(x=factor(`PCOS (Y/N).x`)))+
  geom_bar(stat="count", width=0.7, fill="steelblue")+
  theme_minimal() + labs( x="PCOS(Y/N)", y = "count")
```



```
mean(pcos$`PCOS (Y/N).x`=='1')
```

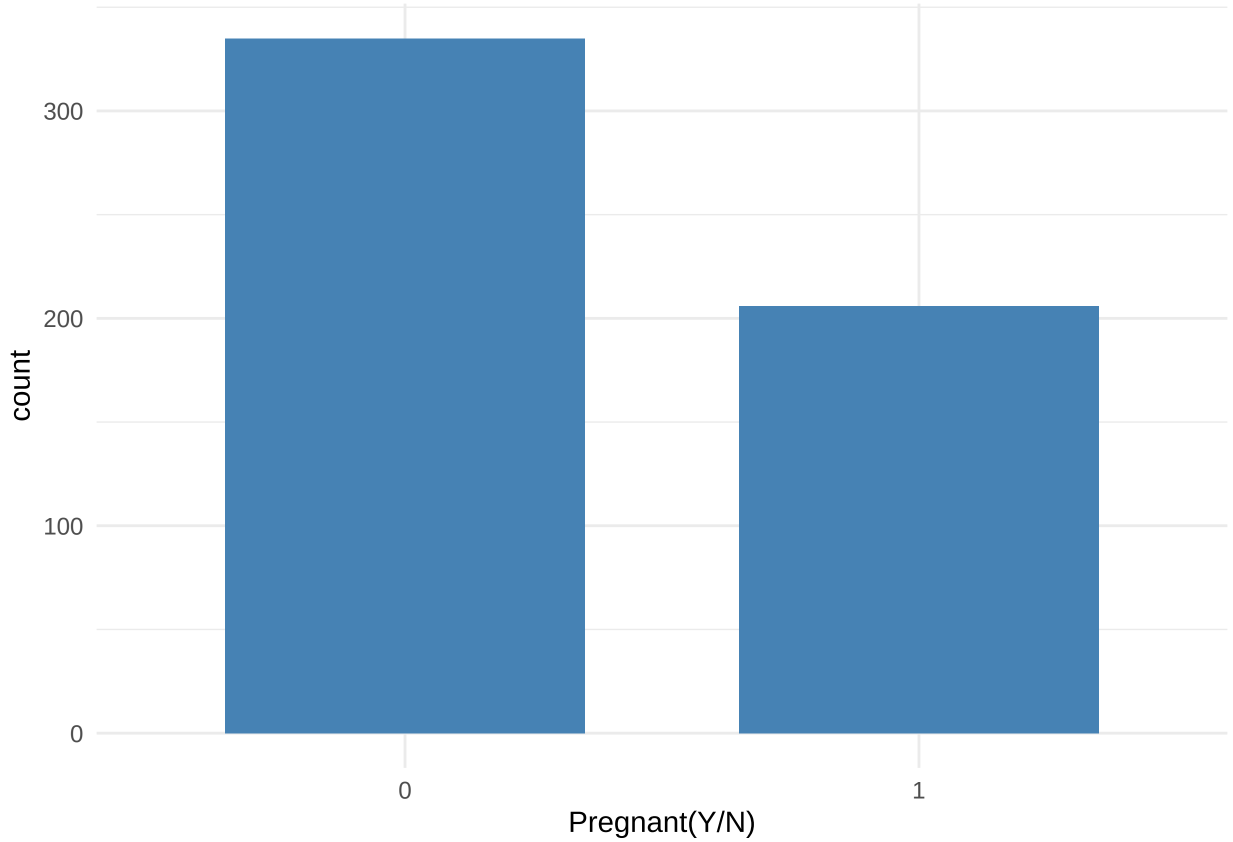
```
## [1] 0.327
```

We can see that proportion of patients with PCOS is 0.327.

Next we will look into the rest of the columns:

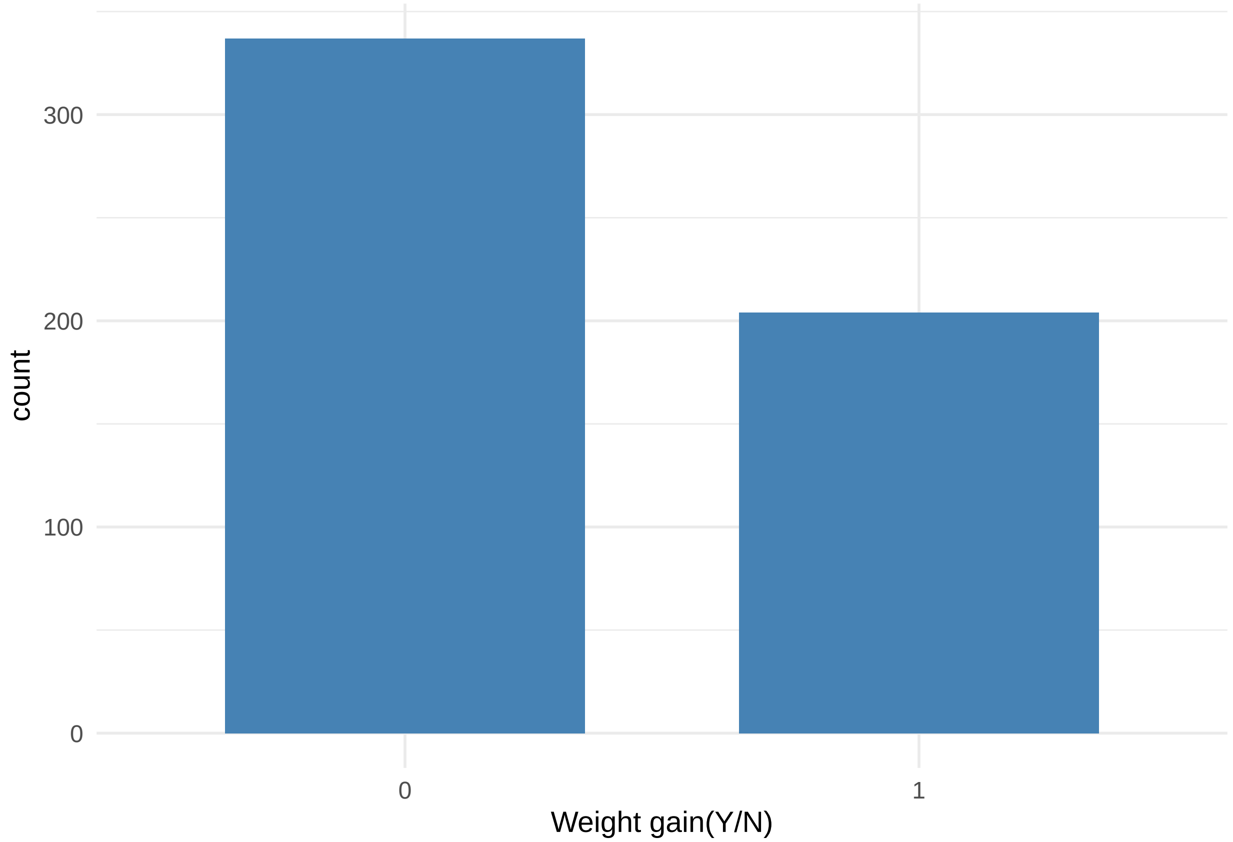
Pregnant(Y/N)

```
ggplot(pcos, aes(x=factor(`Pregnant(Y/N)`)))+
  geom_bar(stat="count", width=0.7, fill="steelblue")+
  theme_minimal() + labs( x="Pregnant(Y/N)", y = "count")
```



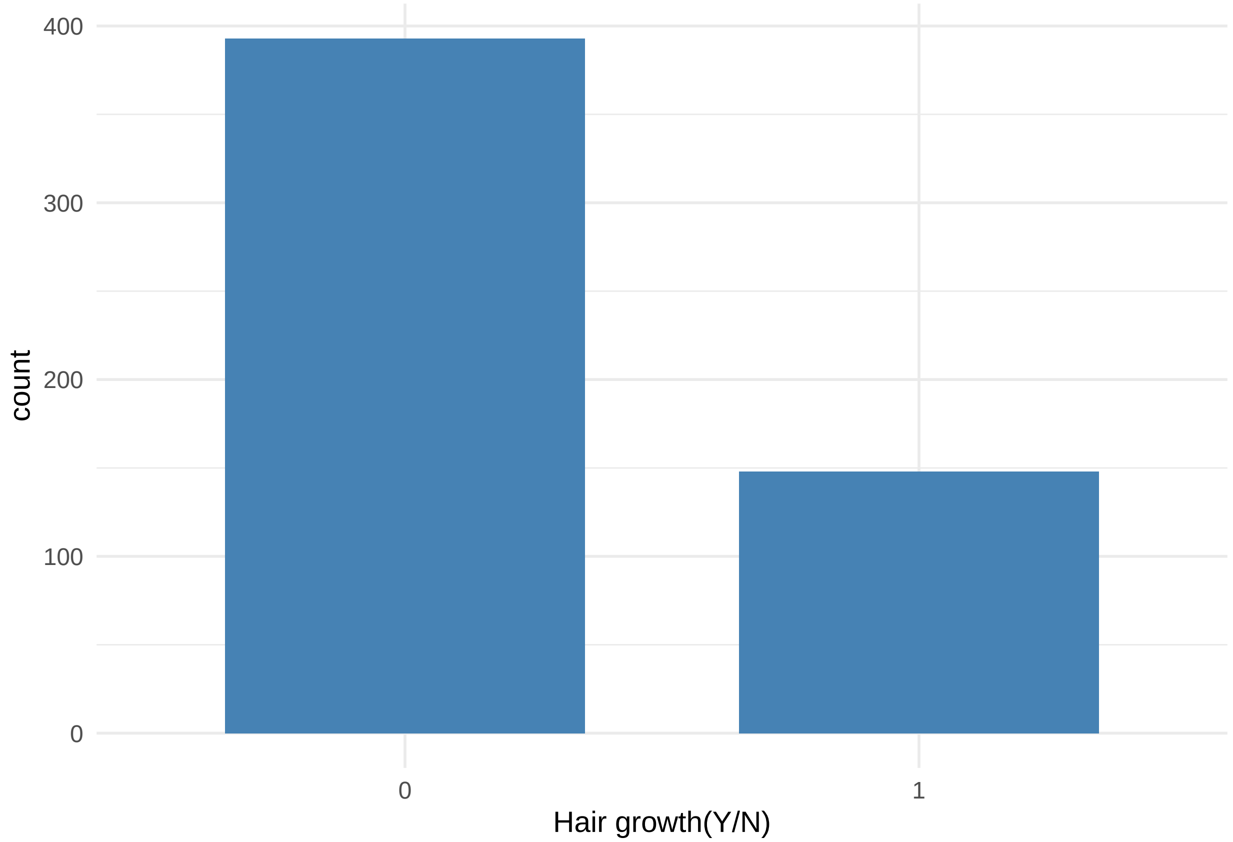
Weight gain(Y/N)

```
ggplot(pcos, aes(x=factor(`Weight gain(Y/N)`)))+
  geom_bar(stat="count", width=0.7, fill="steelblue")+
  theme_minimal() + labs( x="Weight gain(Y/N)", y = "count")
```



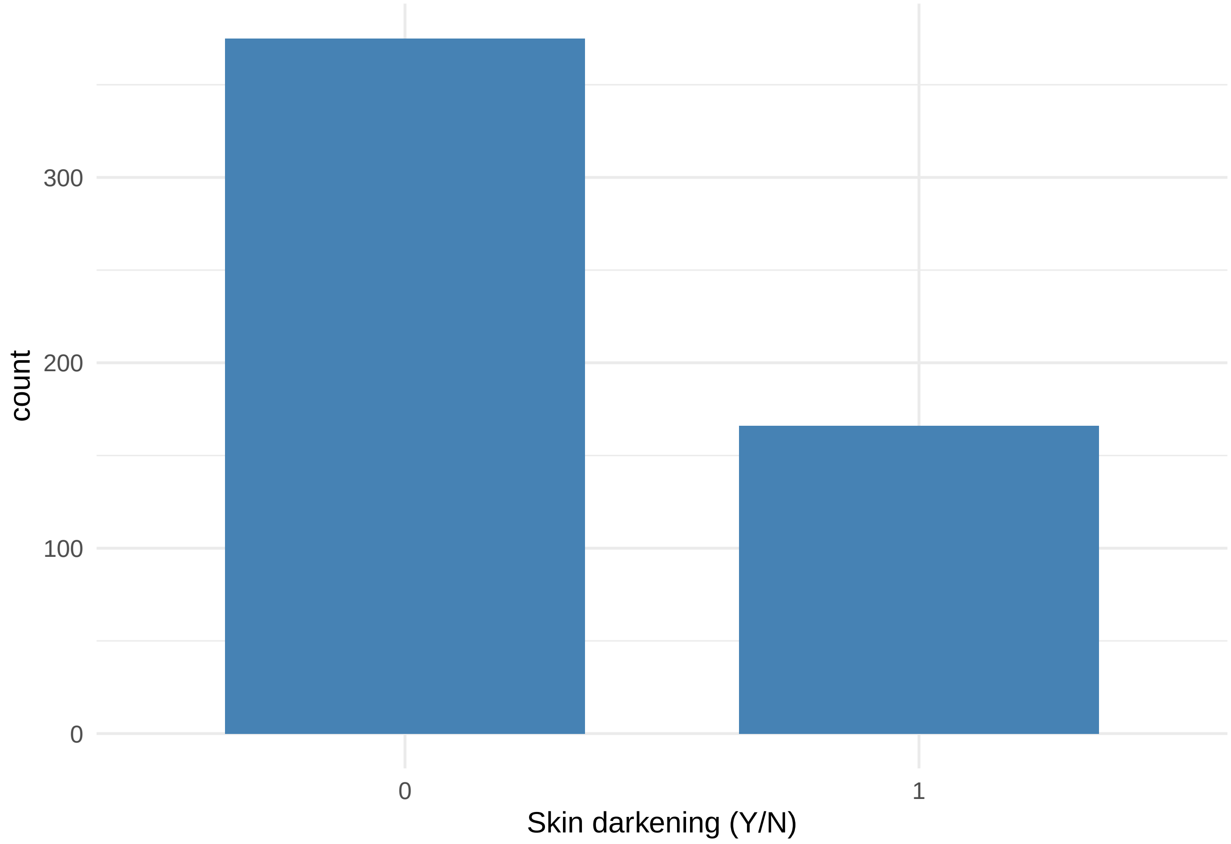
Hair growth(Y/N)

```
ggplot(pcos, aes(x=factor(`hair growth(Y/N)`)))+
  geom_bar(stat="count", width=0.7, fill="steelblue")+
  theme_minimal() + labs( x="Hair growth(Y/N)", y = "count")
```



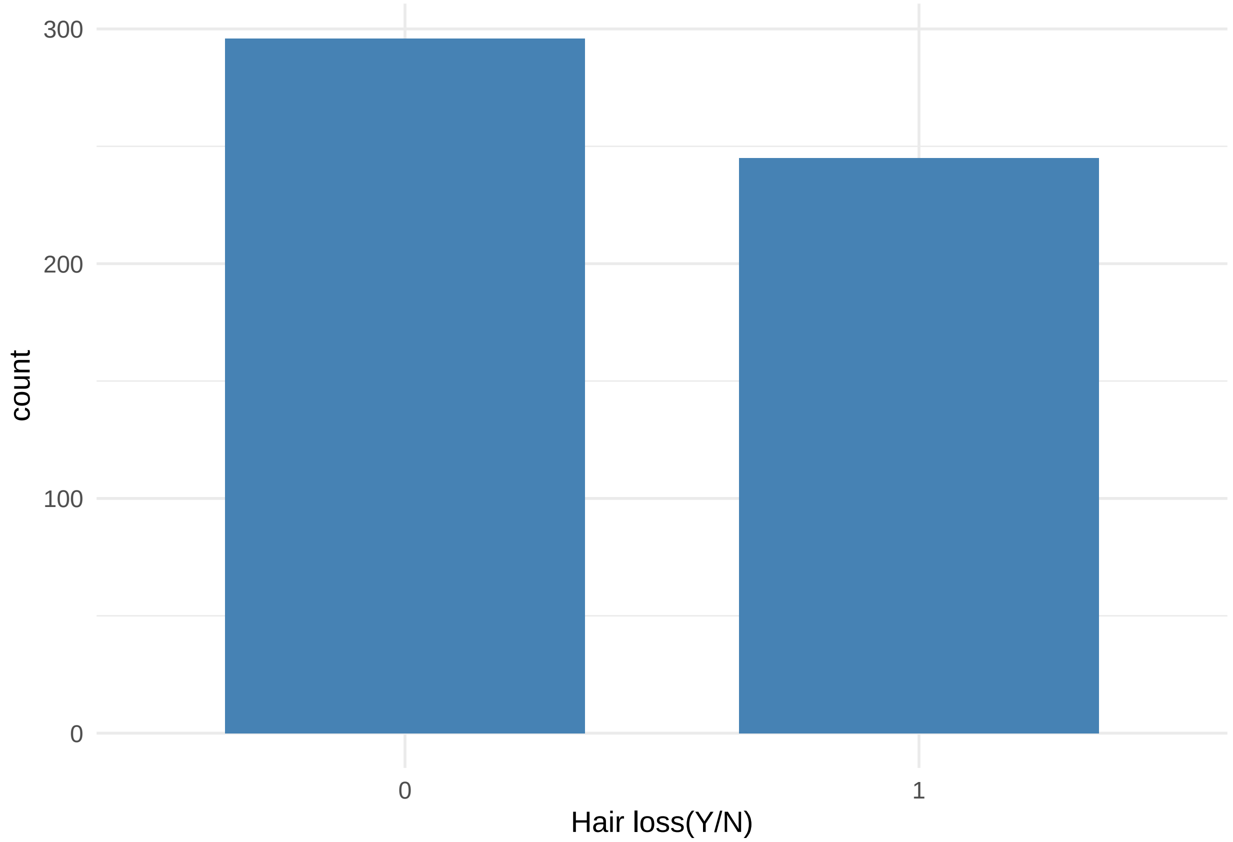
Skin darkening (Y/N)

```
ggplot(pcos, aes(x=factor(`Skin darkening (Y/N)`))+
  geom_bar(stat="count", width=0.7, fill="steelblue")+
  theme_minimal() + labs( x="Skin darkening (Y/N)", y = "count")
```



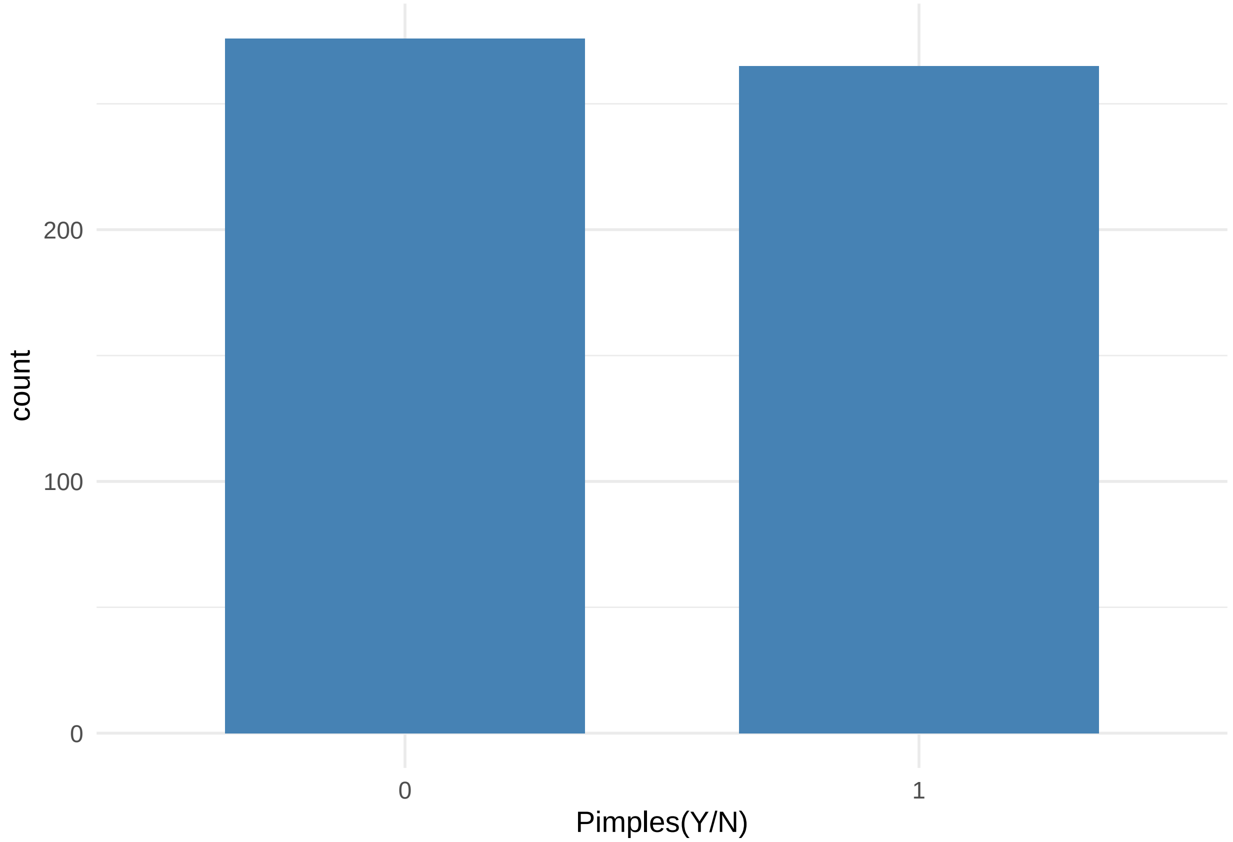
Hair loss(Y/N)

```
ggplot(pcos, aes(x=factor(`Hair loss(Y/N)`)))+
  geom_bar(stat="count", width=0.7, fill="steelblue")+
  theme_minimal() + labs( x="Hair loss(Y/N)", y = "count")
```



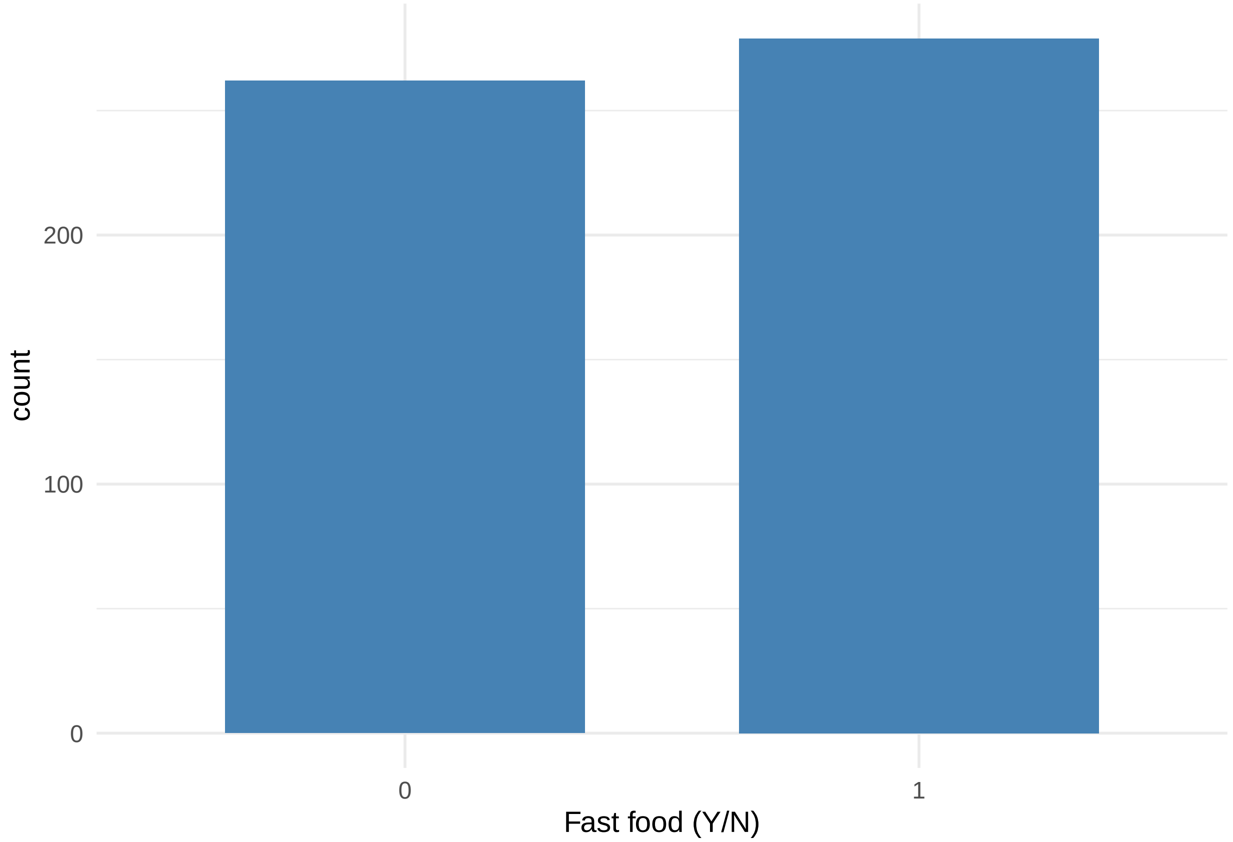
Pimples(Y/N)

```
ggplot(pcos, aes(x=factor(`Pimples(Y/N)`)))+
  geom_bar(stat="count", width=0.7, fill="steelblue")+
  theme_minimal() + labs( x="Pimples(Y/N)", y = "count")
```



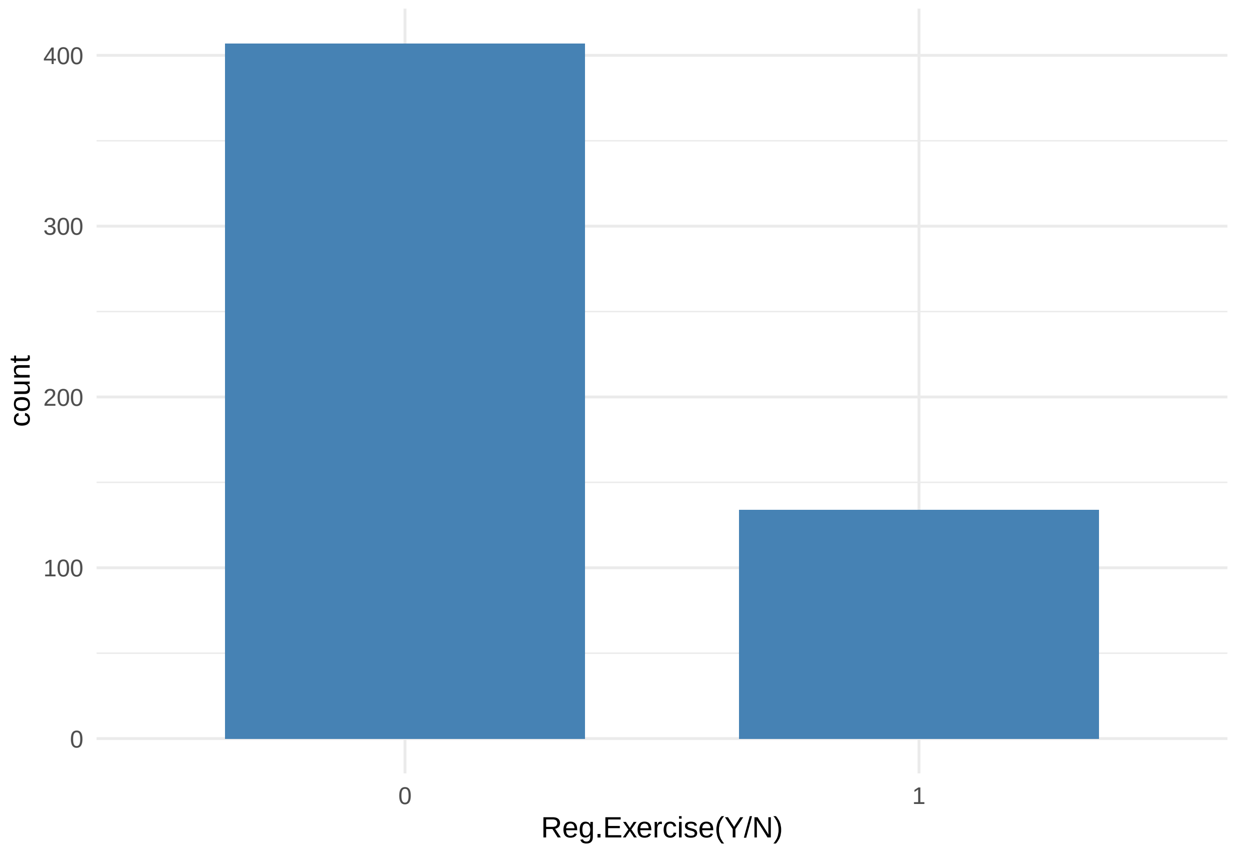
Fast food (Y/N)

```
ggplot(pcos, aes(x=factor(`Fast food (Y/N)`)))+
  geom_bar(stat="count", width=0.7, fill="steelblue")+
  theme_minimal() + labs( x="Fast food (Y/N)", y = "count")
```



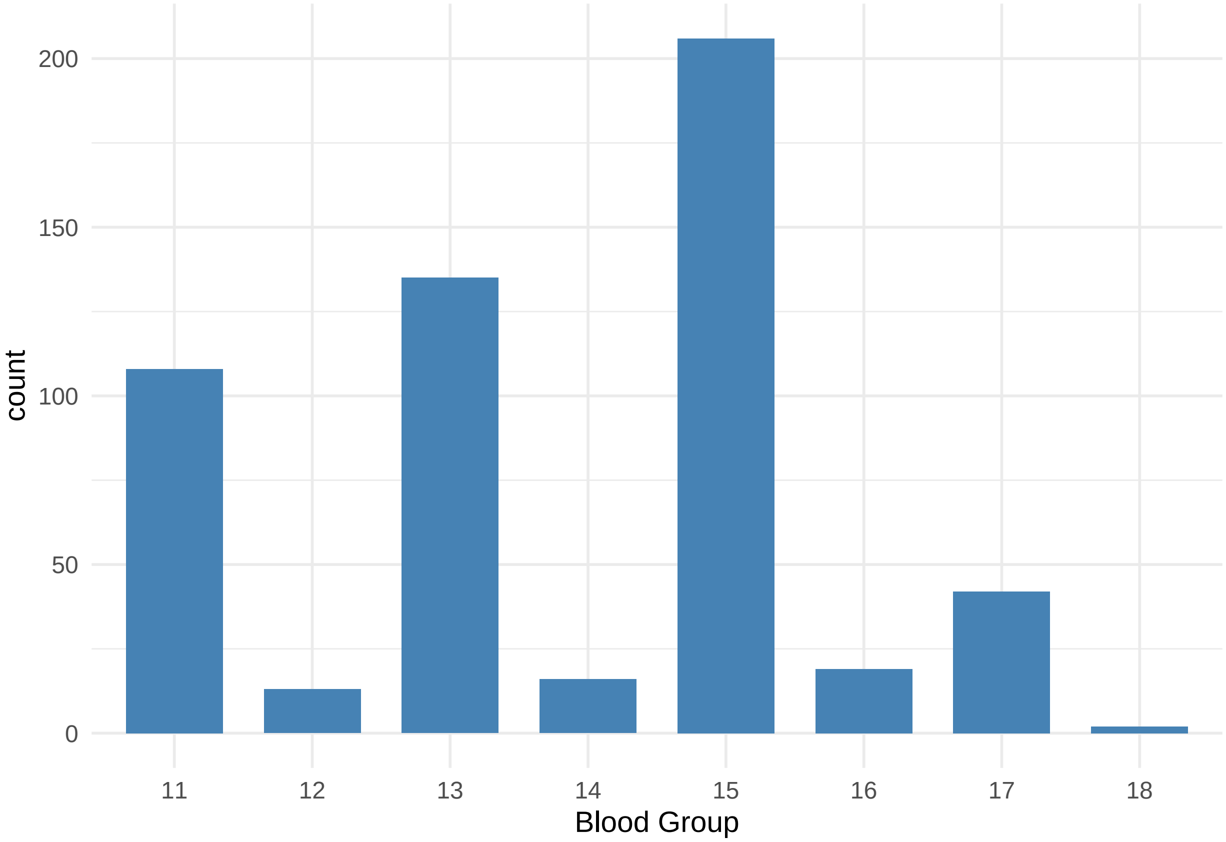
Reg.Exercise(Y/N)

```
ggplot(pcos, aes(x=factor(`Reg.Exercise(Y/N)`)))+
  geom_bar(stat="count", width=0.7, fill="steelblue")+
  theme_minimal() + labs( x="Reg.Exercise(Y/N)", y = "count")
```



Blood Group

```
ggplot(pcos, aes(x=factor(`Blood Group`)))+
  geom_bar(stat="count", width=0.7, fill="steelblue")+
  theme_minimal() + labs( x="Blood Group", y = "count")
```

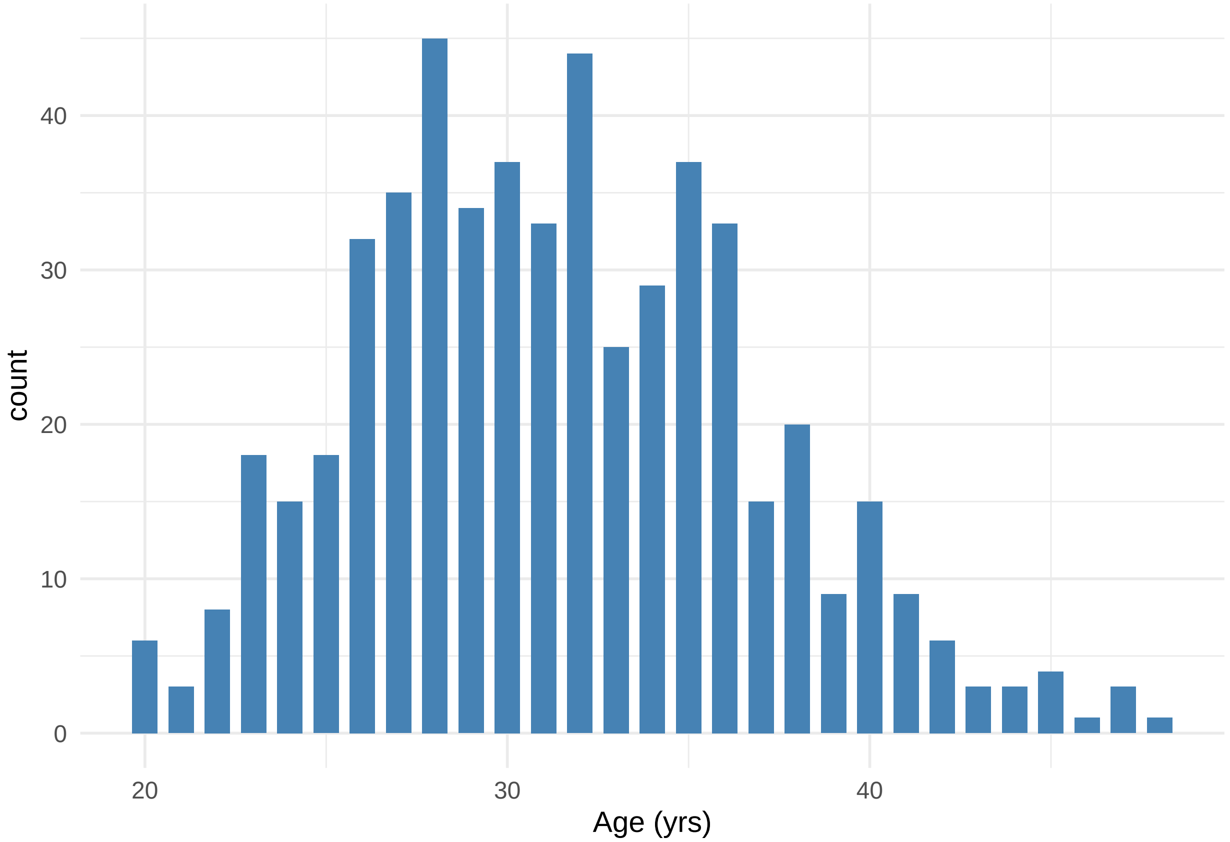


Next we will look into some of the numerical variables. We will check the distribution of Age (yrs), Weight (Kg), BMI, Cycle length(days), Marraige Status (Yrs), No. of aborptions.

Age in yrs

```
ggplot(pcos, aes(x=`Age (yrs)`))+
  geom_histogram(stat="count", width=0.7, fill="steelblue")+
  theme_minimal() + labs( x="Age (yrs)", y = "count")
```

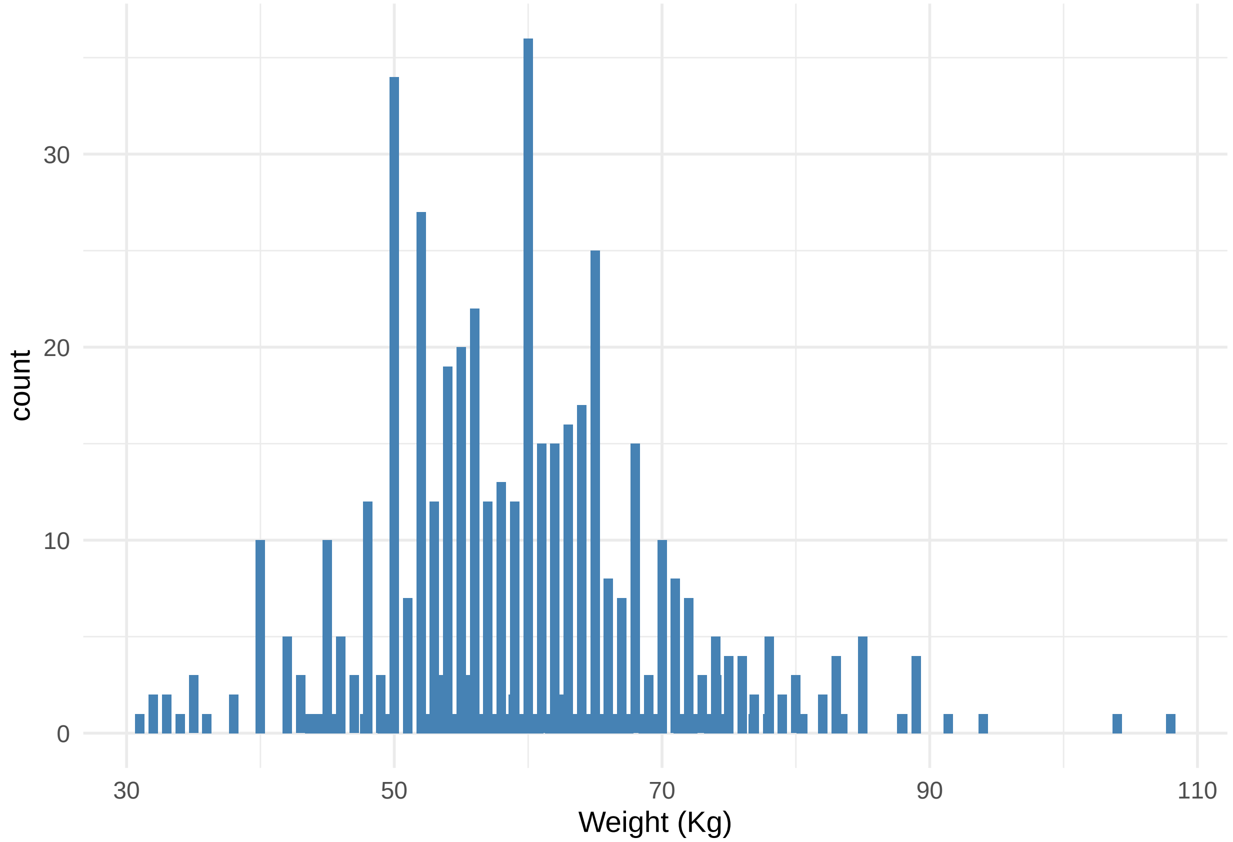
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



Most of the patients is between 25 and 35 years old, which would confirm that from pcos suffer mainly woman in child bearing age.

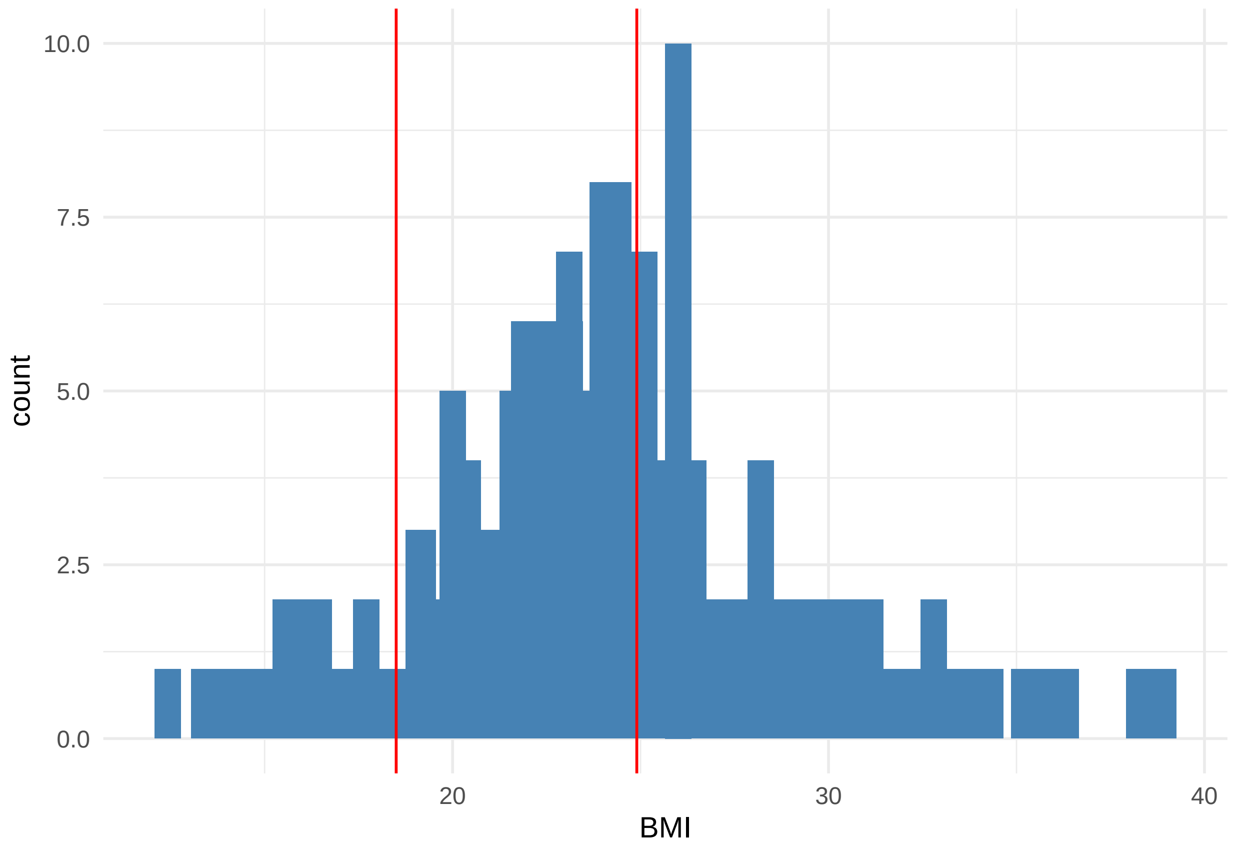
Weight (Kg)

```
ggplot(pcos, aes(x=`Weight (Kg)`))+
  geom_histogram(stat="count", width=0.7, fill="steelblue")+
  theme_minimal() + labs( x="Weight (Kg)", y = "count")
```



BMI

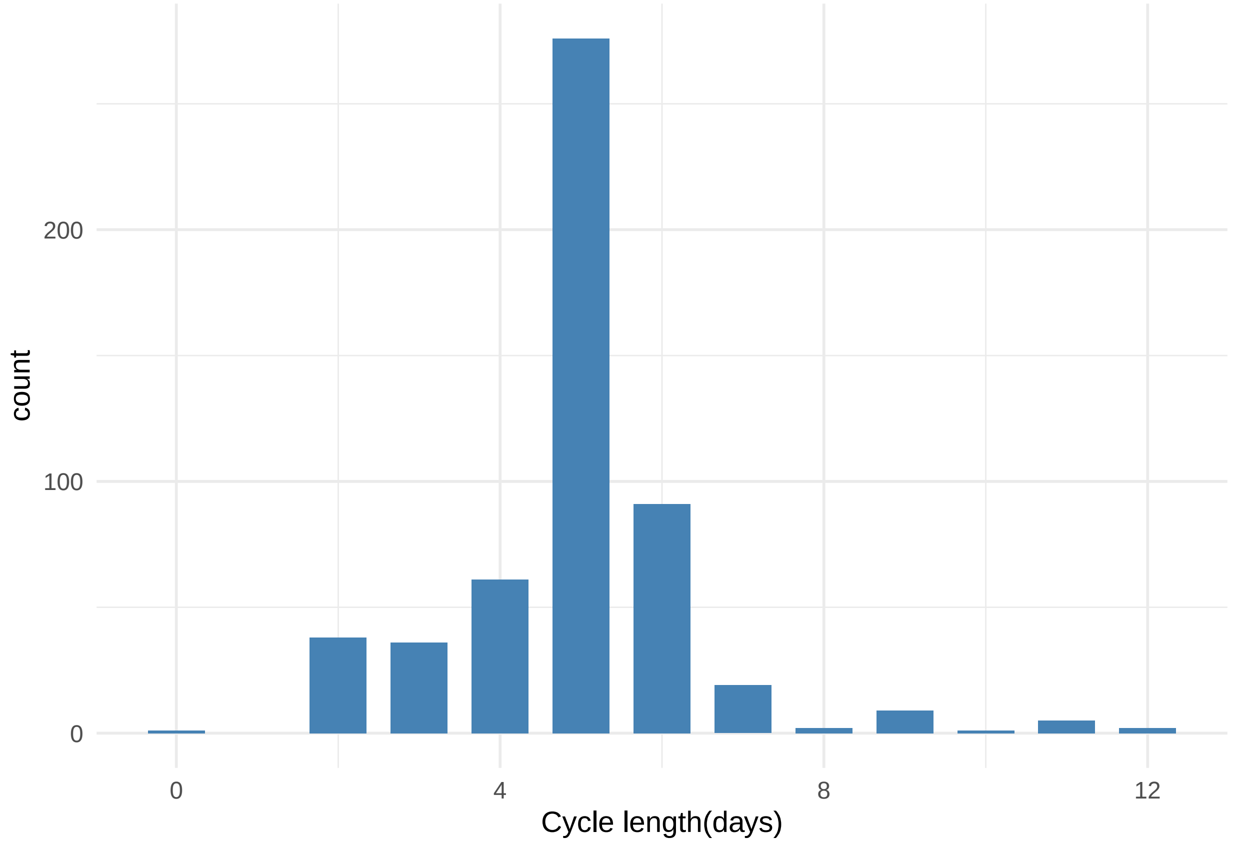
```
ggplot(pcos, aes(x=`BMI`))+
  geom_histogram(stat="count", width=0.7, fill="steelblue")+
  theme_minimal() + labs( x="BMI", y = "count") +
  geom_vline(xintercept = 18.5,color="red") + geom_vline(xintercept = 24.9, color="red")
```



Here we added additional vertical lines, which show the BMI range considered as normal 18.5 - 24.9.

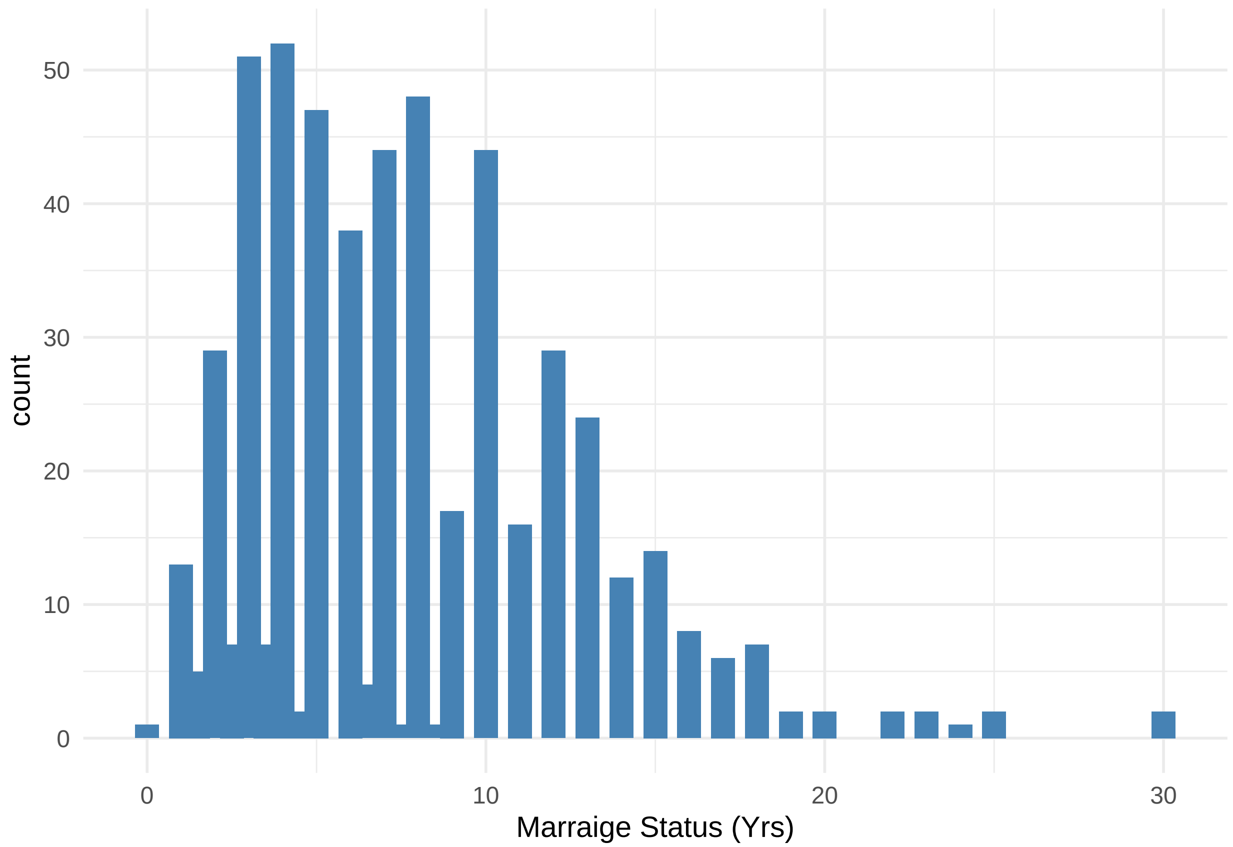
Cycle length(days)

```
ggplot(pcos, aes(x=`Cycle length(days)`))+
  geom_histogram(stat="count", width=0.7, fill="steelblue")+
  theme_minimal() + labs( x="Cycle length(days)", y = "count")
```



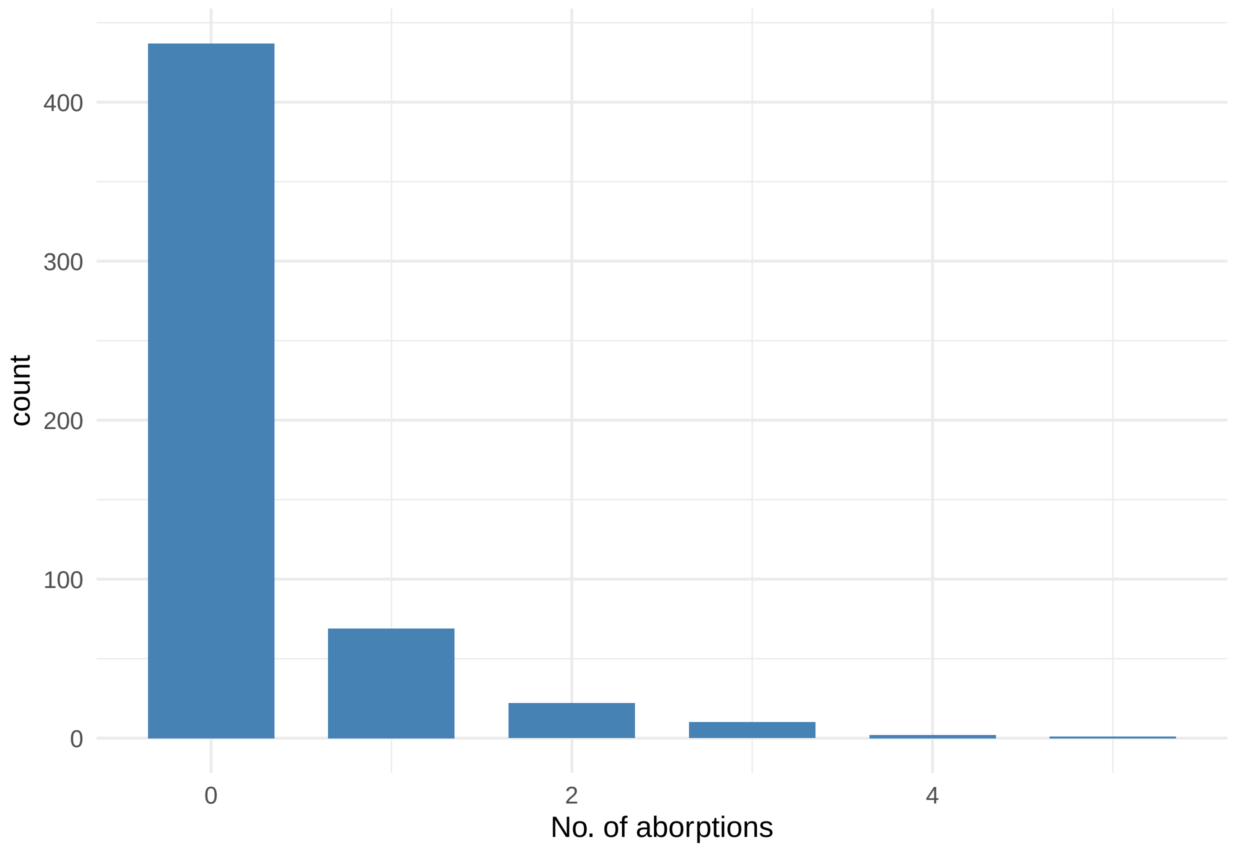
Marraige Status (Yrs)

```
ggplot(pcos, aes(x=`Marraige Status (Yrs)`))+
  geom_histogram(stat="count", width=0.7, fill="steelblue")+
  theme_minimal() + labs( x="Marraige Status (Yrs)", y = "count")
```



No of aborptions

```
ggplot(pcos, aes(x=`No. of aborptions`))+
  geom_histogram(stat="count", width=0.7, fill="steelblue")+
  theme_minimal() + labs( x="No. of aborptions", y = "count")
```



Next we will see the correlation between PCOS (Y/N) column and the rest of the columns. we will concentrate only on significant correlations, where values are above 0.25.

```
cor_pcos <- round(cor(pcos[ , colnames(pcos) != "PCOS (Y/N).x"],
                pcos$`PCOS (Y/N).x`),2)

data.frame(cor_pcos) %>%
  rownames_to_column() %>%
  gather(key="variable", value="correlation", -rowname) %>%
  filter(abs(correlation) > 0.25)
```

```
##                rowname variable correlation
## 1           Cycle(R/I) cor_pcos        0.40
## 2         AMH(ng/mL).x cor_pcos        0.26
## 3      Weight gain(Y/N) cor_pcos        0.44
## 4      hair growth(Y/N) cor_pcos        0.46
## 5 Skin darkening (Y/N) cor_pcos        0.48
## 6         Pimples(Y/N) cor_pcos        0.29
## 7       Fast food (Y/N) cor_pcos        0.38
## 8      Follicle No. (L) cor_pcos        0.60
## 9      Follicle No. (R) cor_pcos        0.65
```

We can see that only 9 variables have a correlation above 0.25.

Fitting a model

Before we fit a model, we need to prepare our data set. That mean we need to divide our target and other variables. Because our target has binary values (0s and 1s), we will change it data type into factor.

```
pcos$`PCOS (Y/N).x` <- factor(pcos$`PCOS (Y/N).x`)

y <- pcos$`PCOS (Y/N).x`

x <-subset(pcos, select = -`PCOS (Y/N).x` )
```

Next we will divide data into train and test set in proportion 20%.

```
set.seed(1)
test_index <- createDataPartition(y, times = 1, p = 0.2, list = FALSE)
test_x <- x[test_index,]
train_x <- x[-test_index,]

test_y <- y[test_index]
train_y <- y[-test_index]
```

Nextly we divide them into train and test sets.

```
set.seed(1)
test_index <- createDataPartition(y, times = 1, p = 0.2, list = FALSE)
test_x <- x[test_index,]
train_x <- x[-test_index,]

test_y <- y[test_index]
train_y <- y[-test_index]
```

Logistic regression

We will first apply the logistic regression.

```
set.seed(1)
train_glm <- train(train_x, train_y, method = "glm")
glm_preds <- predict(train_glm, test_x)
mean(glm_preds == test_y)
```

```
## [1] 0.872
```

The accuracy of this model is 0.872.

Linear discriminant analysis model (LDA)

We can fit the LDA model using caret:

```
set.seed(1)
train_lda <- train(train_x, train_y, method = "lda")
lda_preds <- predict(train_lda, test_x)
mean(lda_preds == test_y)
```

```
## [1] 0.881
```

Not surprisingly the achieved accuracy is similar to the logistic regression, hence the LDA satisfies the assumption of the linear logistic model. If the additional assumption made by LDA is appropriate, LDA tends to estimate the parameters more efficiently by using more information about the data. In practice, logistic regression and LDA often give similar results.

Quadratic discriminant analysis model (QDA)

Lets fit the QDA model with below code:

set.seed(1) train_qda <- train(train_x, train_y, method = “qda”) qda_preds <- predict(train_qda, test_x) mean(qda_preds == test_y)

0.862

With this model we achieve accuracy of 0.862, which is worse than models before.

K-nearest neighbors (kNN)

K-nearest neighbors (kNN) estimates the conditional probabilities in a similar way to bin smoothing. However, kNN is easier to adapt to multiple dimensions. We will try to fit the model with below code:

```
set.seed(1)
train_knn <- train(train_x, train_y,
                   method = "knn")
knn_preds <- predict(train_knn, test_x)
mean(knn_preds == test_y)
```

```
## [1] 0.642
```

The accuracy in this case is 0.642.

K-nearest neighbors (kNN) with cross validation

We will now make a similar analysis but this time we will try to use cross validation to select the optimal k value.

```
set.seed(1)
tuning <- data.frame(k = seq(3, 21, 2))
train_knn_v <- train(train_x, train_y,
                   method = "knn",
                   tuneGrid = tuning)
train_knn_v$bestTune
```

```
##     k
## 10 21
```

```
knn_preds_v <- predict(train_knn_v, test_x)
mean(knn_preds_v == test_y)
```

```
## [1] 0.661
```

The accuracy is only 0.661

Random forest model

Random forests are used in prediction problems where the outcome is categorical - like in our case. The goal is to improve prediction performance and reduce instability by averaging multiple decision trees (a forest of trees constructed with randomness).

```
set.seed(1)
tuning <- data.frame(mtry = c(3, 5, 7, 9))
train_rf <- train(train_x, train_y,
                  method = "rf",
                  tuneGrid = tuning,
                  importance = TRUE)
train_rf$bestTune
```

```
##   mtry
## 4    9
```

```
rf_preds <- predict(train_rf, test_x)
mean(rf_preds == test_y)
```

```
## [1] 0.908
```

The obtained accuracy is 0.908, which is the highest achieved value until now.

We can also see the list of the most important variables in terms of predicting PCOS:

```
varImp(train_rf)
```

```
## rf variable importance
##
##   only 20 most important variables shown (out of 43)
##
##                         Importance
## Follicle No. (R)            100.00
## Follicle No. (L)             57.96
## hair growth(Y/N)             46.34
## Weight gain(Y/N)             39.80
## Skin darkening (Y/N)         35.52
## Cycle(R/I)                   22.92
## Fast food (Y/N)              20.94
## Sl. No                     19.43
## AMH(ng/mL).x               18.18
## Patient File No..x         16.44
## Pimples(Y/N)               16.44
## Cycle length(days)         14.48
## Weight (Kg)                10.83
## Hair loss(Y/N)             10.39
## Hb(g/dl)                    9.74
## BMI                         9.32
## Avg. F size (R) (mm)        8.80
## Reg.Exercise(Y/N)           8.51
## Waist(inch)                 8.35
## Avg. F size (L) (mm)        7.73
```

It looks like the crucial information is the number of follicle in both ovaries (left and right).

Ensembles

The idea of an ensemble is to combine the data from different models to obtain a better estimate.

In machine learning, one can usually greatly improve the final results by combining the results of different algorithms.

ensemble <- cbind(glm = glm_preds == “1”, lda = lda_preds == “1”, qda = qda_preds == “1”, rf = rf_preds == “1”, knn = knn_preds == “1”, knn_v = knn_preds_v == “1”)

ensemble_preds <- ifelse(rowMeans(ensemble) > 0.5, “1”, “0”) mean(ensemble_preds == test_y)

The accuracy in this case is 0.890

Selecting final model

Lets get all the accuracies in one table, to compare which model performs the best in predicting PCOS:

```
##                        Model Accuracy
## 1        Logistic regression    0.872
## 2                        LDA    0.881
## 3                        QDA    0.862
## 4                        Knn    0.642
## 5 Knn with cross validation    0.661
## 6              Random forest    0.908
## 7                   Ensemble    0.890
```

The model with the best accuracy of 0.908 is the random forest model.

Conclusion

Polycystic Ovary Syndrome (PCOS) is a medical condition which causes hormonal disorder in women in their childbearing years. Women with PCOS suffer from many symptoms and risk infertility. The scope of this project was to create a detection system, which would predict whether a woman has or has not PCOS, based on her medical parameters results. The Random Forest Classifier was found to be the most reliable and most accurate among others presented in this paper with accuracy being 90.8%.