Final Project EDA

ALY 6015

Janica St Ville and Jeff Hackmeister

2025-02-24

# Introduction

For this project, we will be working with a dataset we discovered on Kaggle – The “**Polycystic Ovary Syndrome PCOS”.** This dataset speaks about women with Polycystic Ovary Syndrome and is generated according to the Rotterdam criteria. It sparked our interest because it is a prevalent hormonal disorder that can significantly impact women’s health and quality of life.

The goal is to explore how various physiological indicators — such as BMI, hormone levels, and menstrual regularity — relate to the presence of PCOS. I hope to identify patterns that may help in early detection or risk assessment.

To begin the analysis, we will first load in the data from Kaggle and explore the structure of the dataset.

library(tidyverse)  
library(patchwork)  
# read in CSV from https://www.kaggle.com/datasets/lucass0s0/polycystic-ovary-syndrome-pcos  
data <- read.csv("pcos\_rotterdam\_balanceado.csv")  
  
str(data)

'data.frame': 3000 obs. of 6 variables:  
 $ Age : int 29 20 23 19 19 23 21 26 25 31 ...  
 $ BMI : num 21.2 20.5 23.1 32.7 25.9 20.6 25.3 25.2 25.7 29 ...  
 $ Menstrual\_Irregularity : int 0 0 0 1 0 0 0 0 1 0 ...  
 $ Testosterone\_Level.ng.dL.: num 46.1 59.4 69.3 77.7 49.4 36.7 44.2 32.6 96.8 58.3 ...  
 $ Antral\_Follicle\_Count : int 9 6 10 37 5 5 4 4 37 6 ...  
 $ PCOS\_Diagnosis : int 0 0 0 1 0 0 0 0 1 0 ...

The data set contains 3,000 records, 6 variables which are variables of interest including:

Age – Age of the individual

BMI – Body Mass Index

Menstrual\_Irregularity – Indicator of menstrual Irregularities

Testosterone\_Level – measured Testosterone level

Antral\_Follicle\_count – Count of antral follicles

PCOS\_Diagnosis- PCOS diagnosis (likely 0- NO. 1= yes).

## Data Cleaning

First , we will check for missing values.

colSums(is.na(data))

Age BMI Menstrual\_Irregularity   
 0 0 0   
Testosterone\_Level.ng.dL. Antral\_Follicle\_Count PCOS\_Diagnosis   
 0 0 0

To make further analysis easier, we will change the Menstrual\_Irregularity and PCOS\_Diagnosis variables to a factors, as well as shorten the longer variable names.

# Convert to factor with meaningful labels  
data$PCOS\_Diagnosis <- factor(data$PCOS\_Diagnosis, levels = c(0, 1), labels = c("No", "Yes"))  
  
data$Menstrual\_Irregularity <- factor(data$Menstrual\_Irregularity, levels = c(0, 1), labels = c("No", "Yes"))  
  
data <- data %>%  
 rename(T\_Level = Testosterone\_Level.ng.dL.,  
 Men\_Irrg = Menstrual\_Irregularity,  
 AC\_Count = Antral\_Follicle\_Count,  
 PCOS\_diag = PCOS\_Diagnosis)  
  
head(data)

Age BMI Men\_Irrg T\_Level AC\_Count PCOS\_diag  
1 29 21.2 No 46.1 9 No  
2 20 20.5 No 59.4 6 No  
3 23 23.1 No 69.3 10 No  
4 19 32.7 Yes 77.7 37 Yes  
5 19 25.9 No 49.4 5 No  
6 23 20.6 No 36.7 5 No

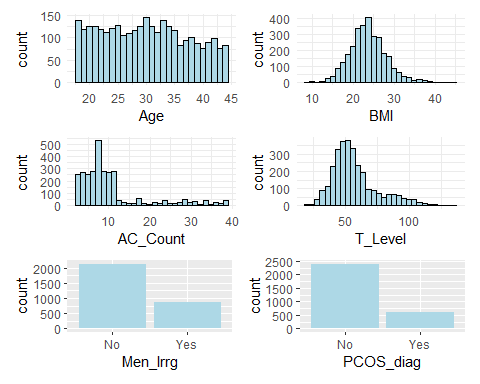
To better compare outcomes further in the analysis, we’ll split the dataset by the diagnosis variable.

y\_diag <- data %>%  
 filter(data$PCOS\_diag == "Yes")  
n\_diag <- data %>%  
 filter(data$PCOS\_diag == "No")

## Variable Distribution

We will be investigating the relationship between the variables and a PCOS diagnosis, it is important to understand the distributions of each variable. W

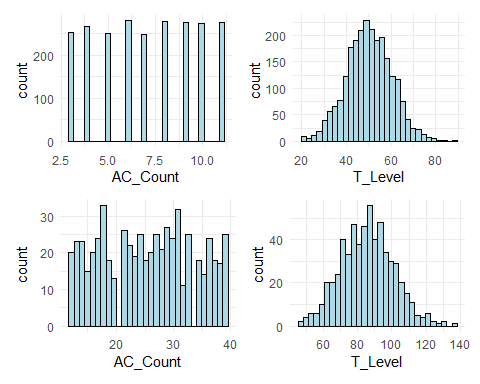
unique\_ages <- seq(min(data$Age) - 0.5, max(data$Age) + 0.5, by = 1)  
age\_plot <- ggplot(data, aes(x=Age)) +   
 geom\_histogram(breaks = unique\_ages,  
 fill = "lightblue",  
 color = "black") +  
 theme\_minimal()  
  
bmi\_plot <- ggplot(data, aes(x = BMI)) +  
 geom\_histogram(fill = "lightblue",  
 color = "black") +  
 theme\_minimal()  
  
mi\_plot <- ggplot(data, aes(x=Men\_Irrg)) +  
 geom\_bar(fill = "lightblue")  
  
t\_plot <- ggplot(data, aes(x = T\_Level)) +  
 geom\_histogram(fill = "lightblue",  
 color = "black") +  
 theme\_minimal()  
  
ac\_plot <- ggplot(data, aes(x = AC\_Count)) +  
 geom\_histogram(fill = "lightblue",  
 color = "black") +  
 theme\_minimal()  
  
pcos\_plot <- ggplot(data, aes(x=PCOS\_diag)) +  
 geom\_bar(fill = "lightblue")  
  
combined\_plot <- (age\_plot + bmi\_plot) / (ac\_plot + t\_plot) / (mi\_plot + pcos\_plot) +   
 plot\_layout(guides = "collect")  
  
print(combined\_plot)

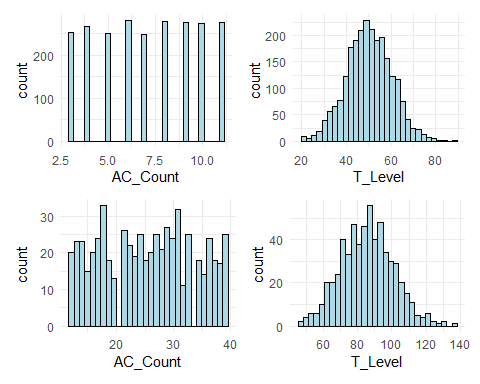


From this we can see a fairly even distribution of ages between 18 and 44, with a slight skew towards the younger ages. The BMI plot produces a mostly normal distribution curve, albeit with a slightly longer tail to the right. The AC count is very left skewed, peaking around 5 and a large drop after 11. Testosterone levels also produced a left skewed distribution with a spike around 50 and a long tale extending well past 100. In the final two plots, you can see the split between the two factor variable.

To take a closer look at the two left skewed variable, we’ll look at their distributions by diagnosis as well.

t\_plot\_y <- ggplot(y\_diag, aes(x = T\_Level)) +  
 geom\_histogram(fill = "lightblue",  
 color = "black") +  
 theme\_minimal()  
  
ac\_plot\_y <- ggplot(y\_diag, aes(x = AC\_Count)) +  
 geom\_histogram(fill = "lightblue",  
 color = "black") +  
 theme\_minimal()  
  
t\_plot\_n <- ggplot(n\_diag, aes(x = T\_Level)) +  
 geom\_histogram(fill = "lightblue",  
 color = "black") +  
 theme\_minimal()  
  
ac\_plot\_n <- ggplot(n\_diag, aes(x = AC\_Count)) +  
 geom\_histogram(fill = "lightblue",  
 color = "black") +  
 theme\_minimal()  
  
(ac\_plot\_n + t\_plot\_n) / (ac\_plot\_y + t\_plot\_y) +   
 plot\_layout(guides = "collect")





# References

Kottarathil, P. (2020, July 11). Polycystic ovary syndrome (PCOS). Kaggle. https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos