Final Project EDA

ALY 6015

Janica St Ville and Jeff Hackmeister

2025-04-26

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. According to the World Health Organization PCOS represents the leading cause of infertility worldwide and nearly 70% of women go undiagnosed. By better understanding the common characteristics of PCOS patients, medical professionals will be better able to target testing efforts and provide treatment. There is no current cure for PCOS, but there are a variety of treatments available to improve symptoms and increase 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
```

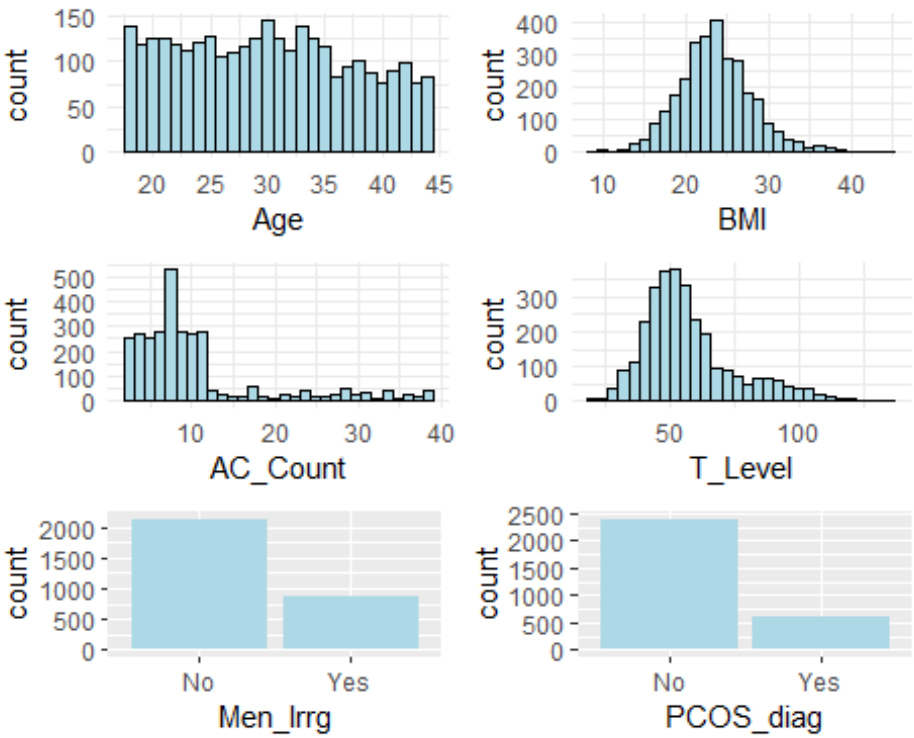
To better compare outcomes further in the analysis, we'll split the dataset by the diagnosis variable.

```
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)
```

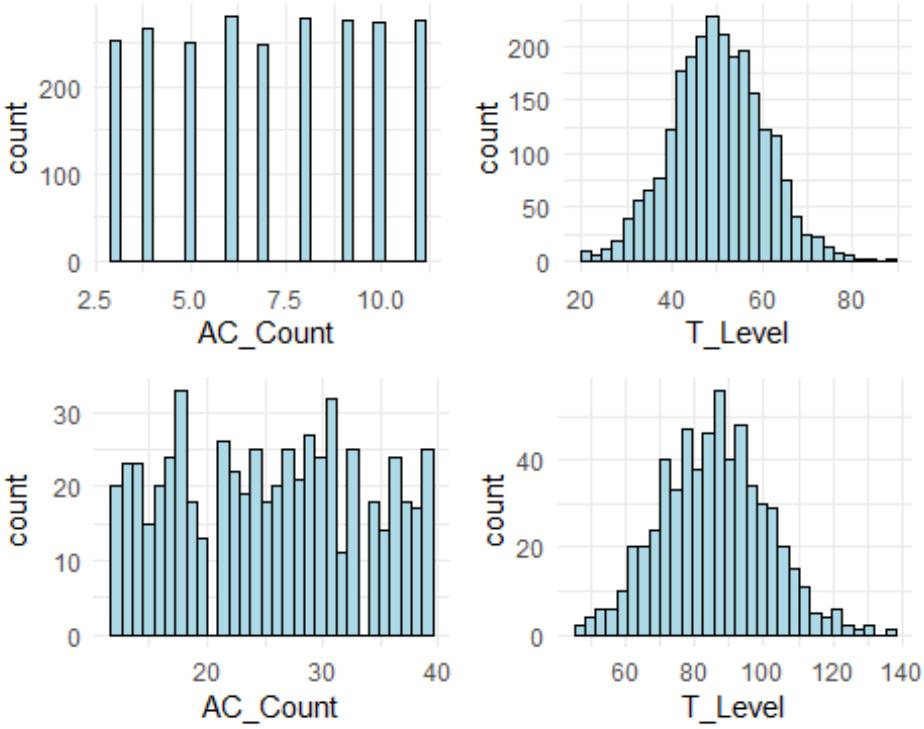


From this we can see a fairly even distribution of ages between 18 and 44, with a slight skew towards the younger ages. The BMI plot produces a mostly normal distribution curve, albeit with a slightly longer tail to the right. The AC count is very left skewed, peaking around 5 and a large drop after 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)) +
```

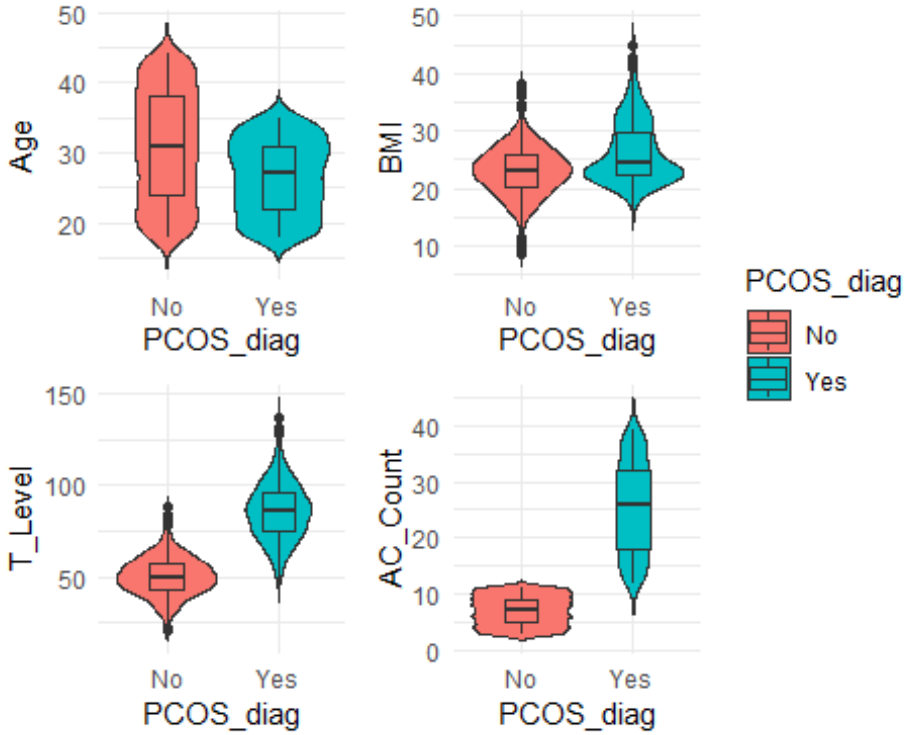
```
  geom_histogram(fill = "lightblue",
                 color = "black") +
  theme_minimal()

(ac_plot_n + t_plot_n) / (ac_plot_y + t_plot_y) +
  plot_layout(guides = "collect")
```



Finally, we'll examine the relationship between each of the numeric variables either a positive or negative PCOS diagnosis. To do this, we'll use the violin plots available in ggplopt2 (part of the tidyverse package), combined with a traditional box plot. This will demonstrate not only the median and quartile information from the box plot, but also the density of the observations at each value.

```
bmi_pcos <- ggplot(data, aes(x = PCOS_diag, y = BMI, fill= PCOS_diag)) +
  geom_violin(trim=FALSE) +
  geom_boxplot(width = 0.3) +
  theme_minimal()

ac_pcos <- ggplot(data, aes(x = PCOS_diag, y = AC_Count, fill= PCOS_diag)) +
  geom_violin(trim=FALSE) +
  geom_boxplot(width = 0.3) +
  theme_minimal()

age_pcos <- ggplot(data, aes(x = PCOS_diag, y = Age, fill= PCOS_diag)) +
  geom_violin(trim=FALSE) +
  geom_boxplot(width = 0.3) +
```

```
  theme_minimal()

t_pcos <- ggplot(data, aes(x = PCOS_diag, y = T_Level, fill= PCOS_diag)) +
  geom_violin(trim=FALSE) +
  geom_boxplot(width = 0.3) +
  theme_minimal()

(age_pcos + bmi_pcos) / (t_pcos + ac_pcos) +
  plot_layout(guides = "collect")
```



From these results, we do not see noticeable differences for age and BMI, but testosterone and especially Antral Follicle values do vary considerably for those with a positive diagnosis. This will certainly be area of further research in this project.

Dataset Analysis and Next Steps

Fortunately, this dataset appears to be fairly clean. We've done some renaming of variables for ease of use but we did not have to work around missing values and there do not appear to be any outliers due to data entry or other errors. This gives us a great starting point for further study. Our next steps should be to further segment the data to look for corralations between natural breaks in the sample population and positive PCOS diagnosis. We will also build a predictive model to determine which combination of factors would lead to a likely positive diagnosis. This model can help direct testing and treatment programs, helping to direct limited resources to likely patients and increase the detection rate.

References

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

World Health Organization. (2025, February 7). *Polycystic ovary syndrome*. World Health Organization; World Health Organization. https://www.who.int/news-room/fact-sheets/detail/polycystic-ovary-syndrome