# Analysis on PCOS Dataset

## Date: - 25th May 2025

## Name: - Sarvesh Sujit Sawant

---

## Introduction: -

Polycystic Ovary Syndrome [PCOS] is a common hormonal disorder in women of reproductive age, often starting during adolescence. Its characterized by hormonal imbalances, irregular periods, excess androgen levels, and cysts in the ovaries. PCOS can impact various aspects of health, including reproductive health, metabolism, and cardiovascular health.

---

## Objective: -

To clean, explore, and derive meaningful insights from PCOS dataset with the goal of understanding influential factors replated to PCOS, and lay the foundation for building predictive models.

---

## Resources: -

Code Editor: - VS Code

Programming Language: - Python

Libraries: - Pandas, NumPy, Matplotlib, Seaborn

---

## Approach: -

Step 1: [Data Cleaning]

- Removing the extra spaces from the column names for proper understanding and readability.
- Removing irrelevant and duplicate columns.
- Handling the missing values using Median/Mode.
- Handling the numeric data in correct format.

Step 2: [Basic Data Exploration]

- Observing the target variable [PCOS (Y/N)].
- Observing the Age Distribution.
- Observing the other parameters like, BMI, AMH, LH, FSH, Follicle Count.

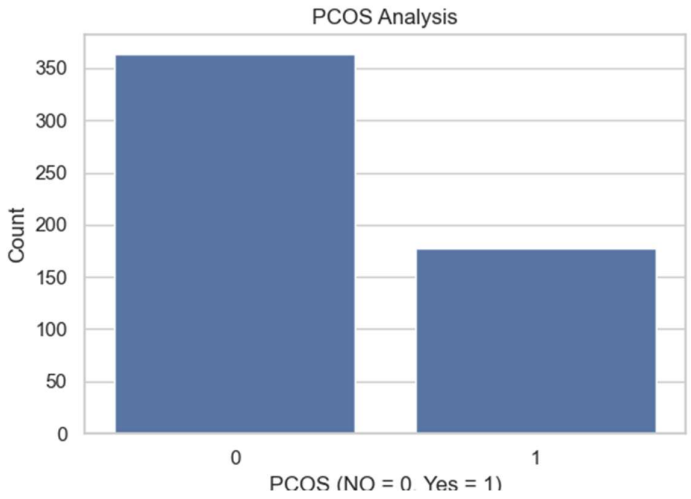
Step 3: [Correlation Analysis]

- Observing the top correlated features with PCOS:
  1. AMH
  2. LH
  3. Follicle Count
  4. BMI
  5. Weight

---

## Result: -

1. Target Variable Distribution [PCOS (Y/N)]:



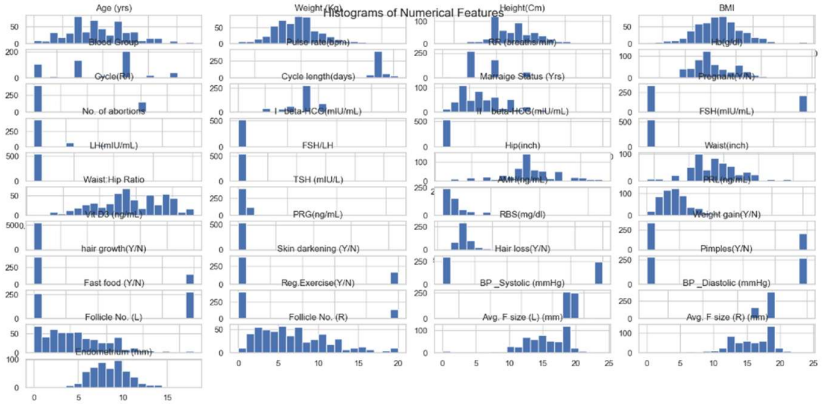

Observation: -

- [0 (No PCOS)] cases are more frequent
- [1 (PCOS)] cases are less frequent

Insights: -

- The dataset is imbalanced
- The training model may become biased to the majority

2. Histogram Analysis of Numerical Features:



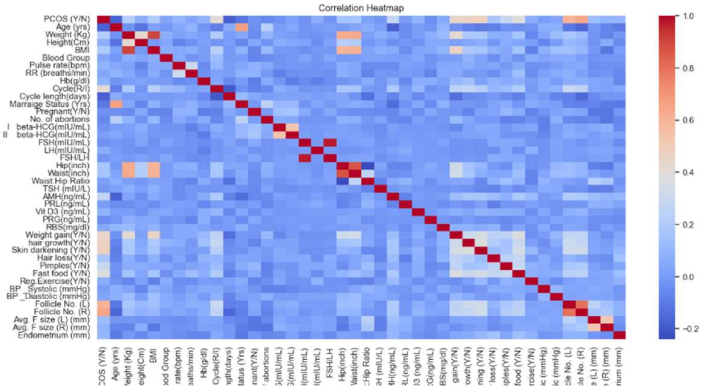

Observation: -

- Features like “Weight”, “BMI”, “AMH(ng/mL)”, “TSH” are right skewed with outliers.
- Features like “Age” and “Pulse rate” are normally distributed.

Insights: -

- Right-Skewed features may benefit from log transformation.
- It helps to decide which features might need standardization or normalization.

3. Correlation Analysis:



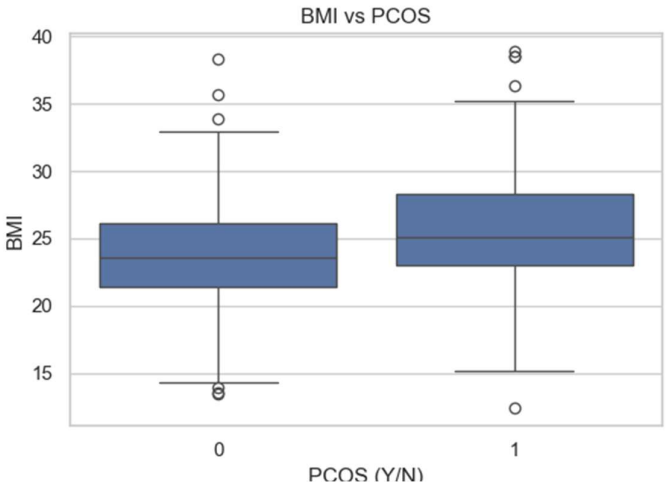

Observation: -

- “Weight” and “BMI” are strongly positively correlated.
- “LH” and “FSH” show moderate correlation.
- “AMH(ng/mL)” is strongly correlated with [PCOS (Y/N)].

Insights: -

- Features like “AMH”, “LH”, “FSH” might be strong prdictors of PCOS.
- Some features like “BP”, “Pulse rate” have weak correlation with PCOS.
- “BMI” and “Weight” are redundant features.

4. Boxplot Analysis:



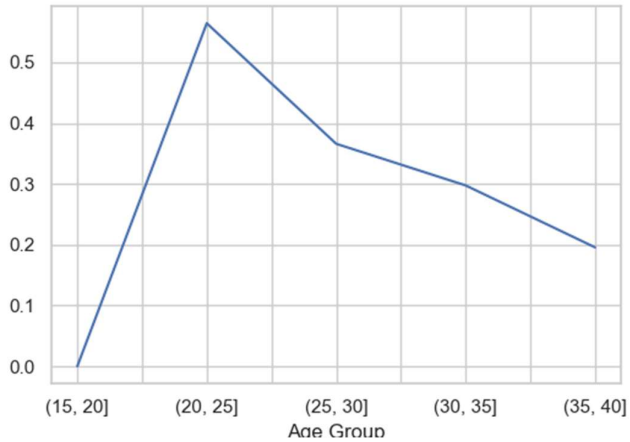

Observation: -

- [0 (No PCOS)] group has low level of BMI.
- [1 (PCOS)] group has comparatively higher level of BMI.

Insights: -

- The difference in BMI level helps in classification of PCOS group.

5. Age Analysis: [EXTRA]




Observation: -

- Peak in the age group of 20 to 25.

Insights: -

- PCOS is highly prominent at the age of 20 to 25.

---

## Conclusion: -

This detailed EDA not only validates known medical correlations with PCOS but also uncovers deeper patterns through clustering and feature engineering. The next steps include predictive model building using these insights and deploying explainable ML systems to assist healthcare professionals.