PCOD DATASET ANALYSIS

MOULIKA KOLAVASI

INTRODUCTION

Polycystic ovary syndrome (PCOS) is one of the most common yet underdiagnosed hormonal disorders in women, often affecting overall health, fertility, and quality of life.

Early detection is crucial, but symptoms are often overlooked or misunderstood. By analyzing real-world health data, this project aims to uncover hidden patterns and risk factors that could support early diagnosis and better awareness.

With the growing focus on women's health, using data analytics to explore PCOS is both relevant and impactful.

TERMS USED IN THIS PROJECT FOR REFERENCE

- **BMI** – *Body Mass Index*, used to assess body fat based on height and weight.
- **LH** – *Luteinizing Hormone*, triggers ovulation.
- **FSH** – *Follicle-Stimulating Hormone*, helps regulate the menstrual cycle.
- **PRL** – *Prolactin*, controls milk production and affects fertility.
- **TSH** – *Thyroid-Stimulating Hormone*, regulates thyroid function and metabolism.

Objectives

- Analyze patterns and correlations in PCOS-related health data
- Understand the impact of clinical factors on PCOS diagnosis
- Tools used: Excel and Power Query

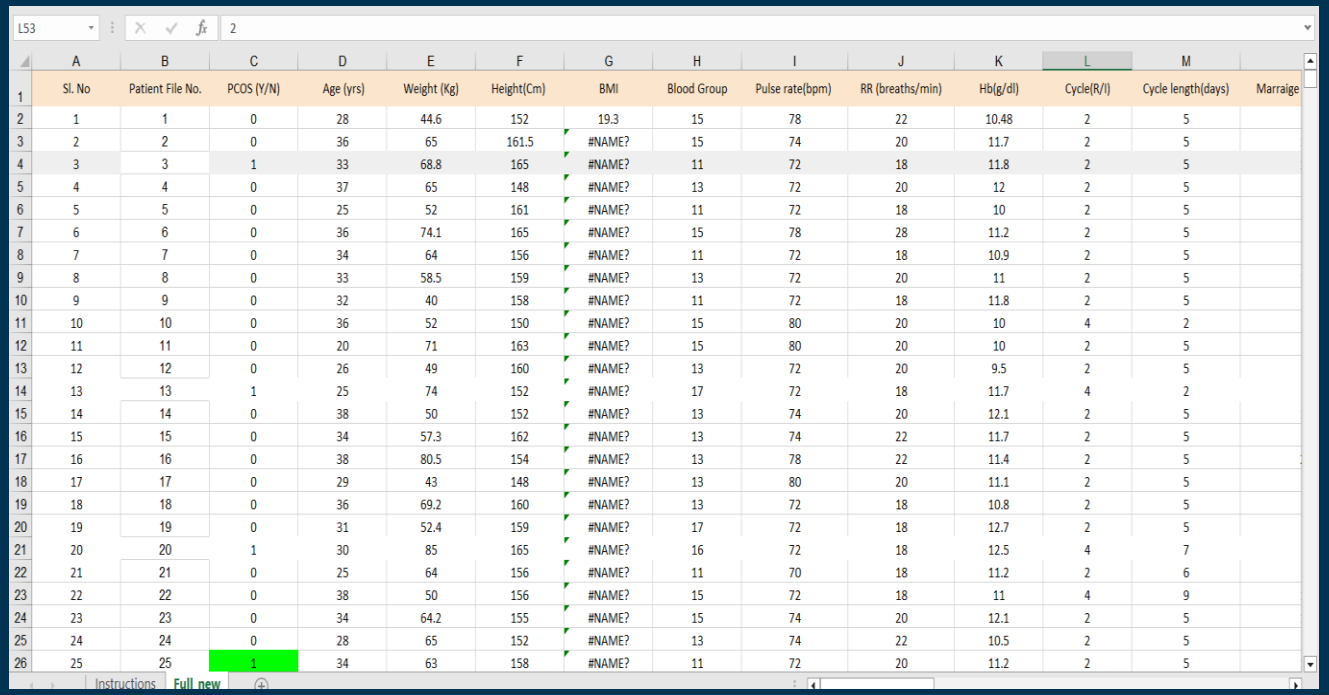
RAW Vs CLEANED DATA TABLES

Issues

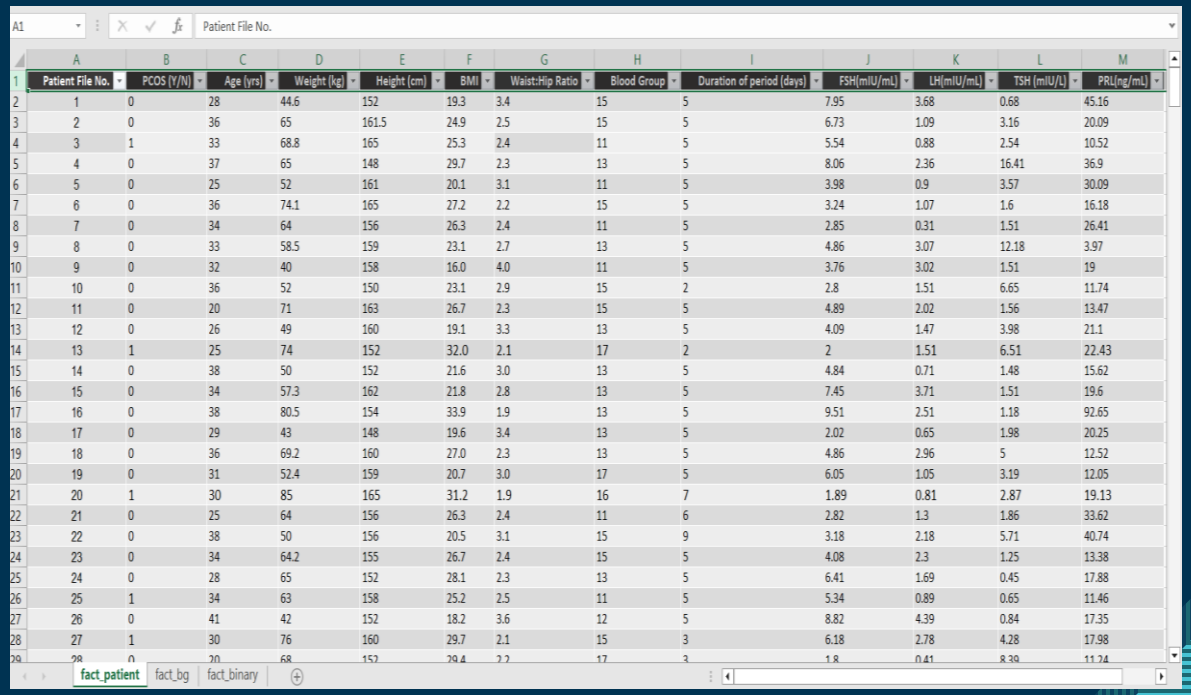
- Null data points
- Inconsistencies
- Extra columns

Rectified

- ✓ Nulls removed
- ✓ Column names standardized
- ✓ Irrelevant data dropped

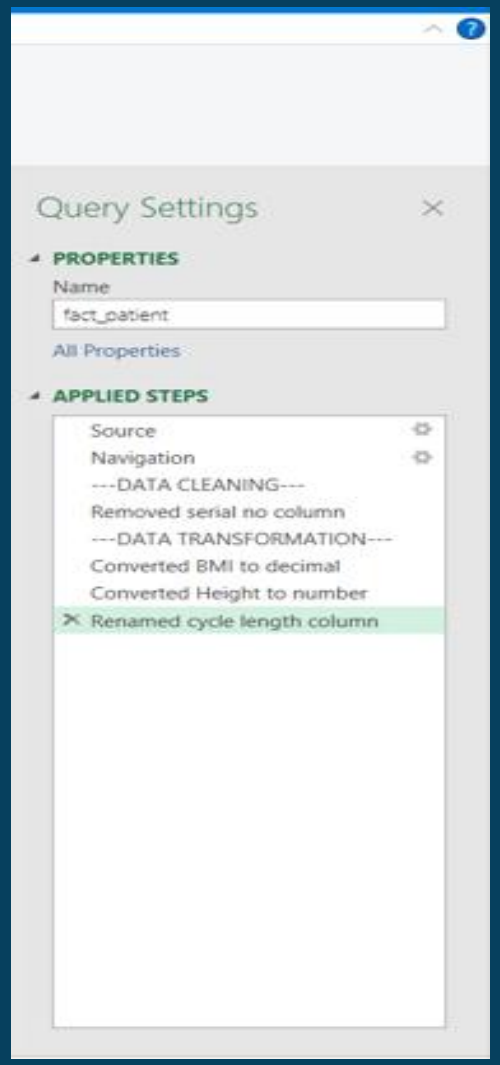


Sl. No	Patient File No.	PCOS (Y/N)	Age (yrs)	Weight (Kg)	Height(Cm)	BMI	Blood Group	Pulse rate(bpm)	RR (breaths/min)	Hb(g/dl)	Cycle(R/I)	Cycle length(days)	Marraige
1	1	0	28	44.6	152	19.3	15	78	22	10.48	2	5	
2	2	0	36	65	161.5	#NAME?	15	74	20	11.7	2	5	
3	3	1	33	68.8	165	#NAME?	11	72	18	11.8	2	5	
4	4	0	37	65	148	#NAME?	13	72	20	12	2	5	
5	5	0	25	52	161	#NAME?	11	72	18	10	2	5	
6	6	0	36	74.1	165	#NAME?	15	78	28	11.2	2	5	
7	7	0	34	64	156	#NAME?	11	72	18	10.9	2	5	
8	8	0	33	58.5	159	#NAME?	13	72	20	11	2	5	
9	9	0	32	40	158	#NAME?	11	72	18	11.8	2	5	
10	10	0	36	52	150	#NAME?	15	80	20	10	4	2	
11	11	0	20	71	163	#NAME?	15	80	20	10	2	5	
12	12	0	26	49	160	#NAME?	13	72	20	9.5	2	5	
13	13	1	25	74	152	#NAME?	17	72	18	11.7	4	2	
14	14	0	38	50	152	#NAME?	13	74	20	12.1	2	5	
15	15	0	34	57.3	162	#NAME?	13	74	22	11.7	2	5	
16	16	0	38	80.5	154	#NAME?	13	78	22	11.4	2	5	
17	17	0	29	43	148	#NAME?	13	80	20	11.1	2	5	
18	18	0	36	69.2	160	#NAME?	13	72	18	10.8	2	5	
19	19	0	31	52.4	159	#NAME?	17	72	18	12.7	2	5	
20	20	1	30	85	165	#NAME?	16	72	18	12.5	4	7	
21	21	0	25	64	156	#NAME?	11	70	18	11.2	2	6	
22	22	0	38	50	156	#NAME?	15	72	18	11	4	9	
23	23	0	34	64.2	155	#NAME?	15	74	20	12.1	2	5	
24	24	0	28	65	152	#NAME?	13	74	22	10.5	2	5	
25	25	1	34	63	158	#NAME?	11	72	20	11.2	2	5	



Patient File No.	PCOS (Y/N)	Age (yrs)	Weight (kg)	Height (cm)	BMI	Waist:Hip Ratio	Blood Group	Duration of period (days)	FSH(mIU/mL)	LH(mIU/mL)	TSH (mIU/L)	PRL(ng/mL)
1	0	28	44.6	152	19.3	3.4	15	5	7.95	3.68	0.68	45.16
2	0	36	65	161.5	24.9	2.5	15	5	6.73	1.09	3.16	20.09
3	1	33	68.8	165	25.3	2.4	11	5	5.54	0.88	2.54	10.52
4	0	37	65	148	29.7	2.3	13	5	8.06	2.36	16.41	36.9
5	0	25	52	161	20.1	3.1	11	5	3.98	0.9	3.57	30.09
6	0	36	74.1	165	27.2	2.2	15	5	3.24	1.07	1.6	16.18
7	0	34	64	156	26.3	2.4	11	5	2.85	0.31	1.51	26.41
8	0	33	58.5	159	23.1	2.7	13	5	4.86	3.07	12.18	3.97
9	0	32	40	158	16.0	4.0	11	5	3.76	3.02	1.51	19
10	0	36	52	150	23.1	2.9	15	2	2.8	1.51	6.65	11.74
11	0	20	71	163	26.7	2.3	15	5	4.89	2.02	1.56	13.47
12	0	26	49	160	19.1	3.3	13	5	4.09	1.47	3.98	21.1
13	1	25	74	152	32.0	2.1	17	2	2	1.51	6.51	22.43
14	0	38	50	152	21.6	3.0	13	5	4.84	0.71	1.48	15.62
15	0	34	57.3	162	21.8	2.8	13	5	7.45	3.71	1.51	19.6
16	0	38	80.5	154	33.9	1.9	13	5	9.51	2.51	1.18	92.65
17	0	29	43	148	19.6	3.4	13	5	2.02	0.65	1.98	20.25
18	0	36	69.2	160	27.0	2.3	13	5	4.86	2.96	5	12.52
19	0	31	52.4	159	20.7	3.0	17	5	6.05	1.05	3.19	12.05
20	1	30	85	165	31.2	1.9	16	7	1.89	0.81	2.87	19.13
21	0	25	64	156	26.3	2.4	11	6	2.82	1.3	1.86	33.62
22	0	38	50	156	20.5	3.1	15	9	3.18	2.18	5.71	40.74
23	0	34	64.2	155	26.7	2.4	15	5	4.08	2.3	1.25	13.38
24	0	28	65	152	28.1	2.3	13	5	6.41	1.69	0.45	17.88
25	1	34	63	158	25.2	2.5	11	5	5.34	0.89	0.65	11.46
26	0	41	42	152	18.2	3.6	12	5	8.82	4.39	0.84	17.35
27	1	30	76	160	29.7	2.1	15	3	6.18	2.78	4.28	17.98

ETL (Extract, Transform, Load)



Further data cleaning and transformation were performed by extracting and loading the data into the Power Query Editor

Data Cleaning –

- Duplicate and redundant columns were

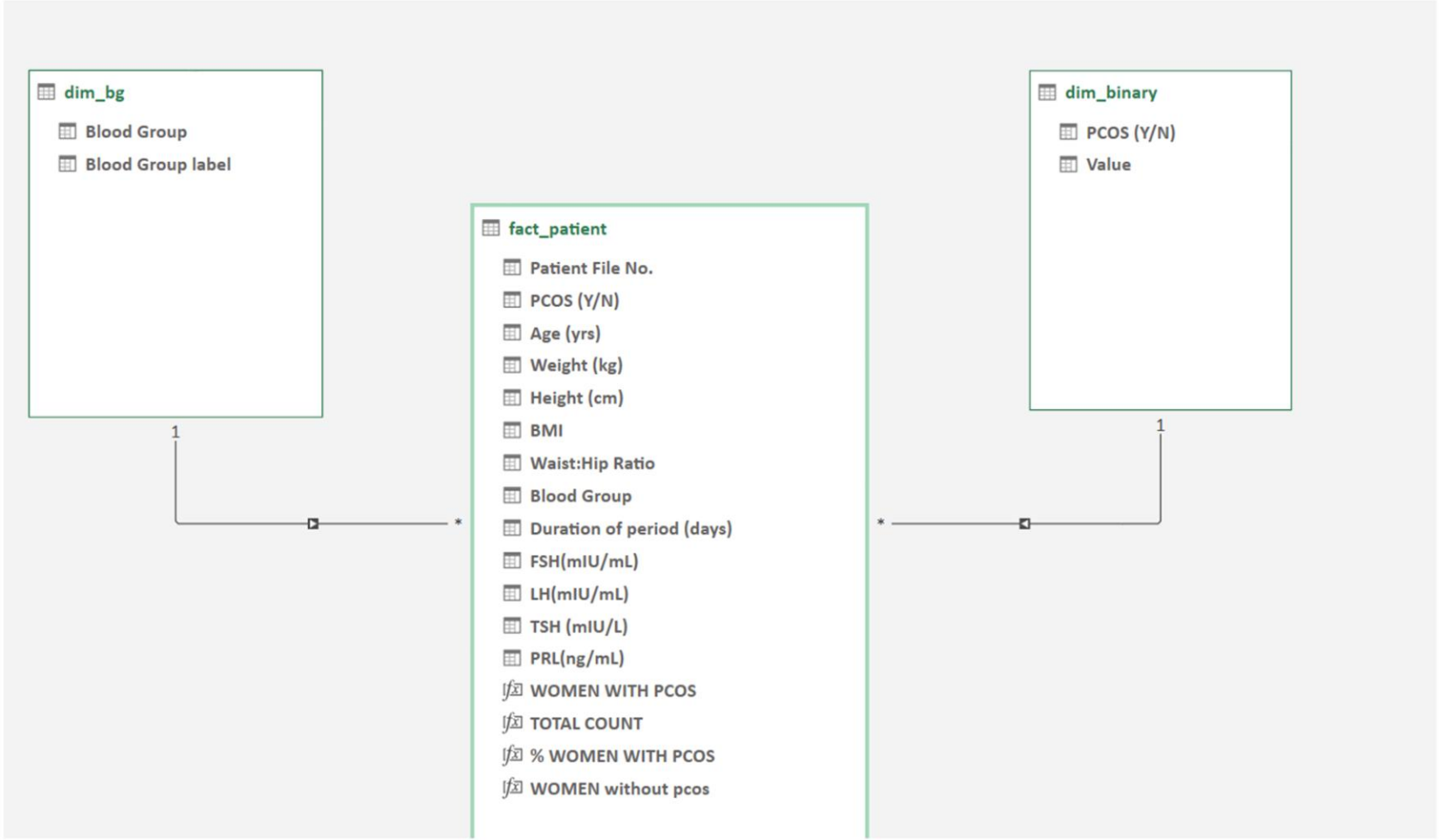
Removed

Data Transformation –

- Column data types were adjusted for better relevance (e.g., BMI rounded to 2 decimal places, height converted to a whole number).
- Columns were renamed for clarity (e.g., *Cycle Length* renamed to *Duration of Period (days)*).

The data was then loaded back to the worksheets and added to the data model

DATA MODELLING



DATA MODEL DIAGRAM

- The data was prepared for data modeling
- A data model was created by linking appropriate columns from the dimension tables to the fact tables.
- A one-to-many relationship was established.
- Data modeling was performed using the star schema technique

PIVOT TABLES AND CONDITIONAL FORMATTING

WOMEN DIAGNOSED AND NOT DIAGNOSED

Status	Count	Percentage
Diagnosed	125	31.3%
Not Diagnosed	275	68.8%
Grand Total	400	

WAIST : HIP RATIO OF DIAGNOSED AND NOT DIAGNOSED

Row Labels	Average of Waist:Hip Ratio
Diagnosed	2.562174158
Not Diagnosed	2.766210828

DIAGNOSED AND NOT DIAGNOSED : AVERAGE HORMONE LEVELS

Row Labels	FSH(mIU/mL)	TSH (mIU/L)	LH(mIU/mL)	PRL(ng/mL)
Diagnosed	5.15	3.12	3.12	24.26
Not Diagnosed	23.66	3.12	2.61	24.11

BLOOD GROUPS : DIAGNOSED AND % OF TOTAL DIAGNOSED

Row Labels	Women with PCOS	Percentage
A−	1	0.8%
A+	22	17.6%
AB-	1	0.8%
AB+	15	12.0%
B−	5	4.0%
B+	24	19.2%
O-	6	4.8%
O+	51	40.8%
Grand Total	125	

AGE : DIAGNOSED

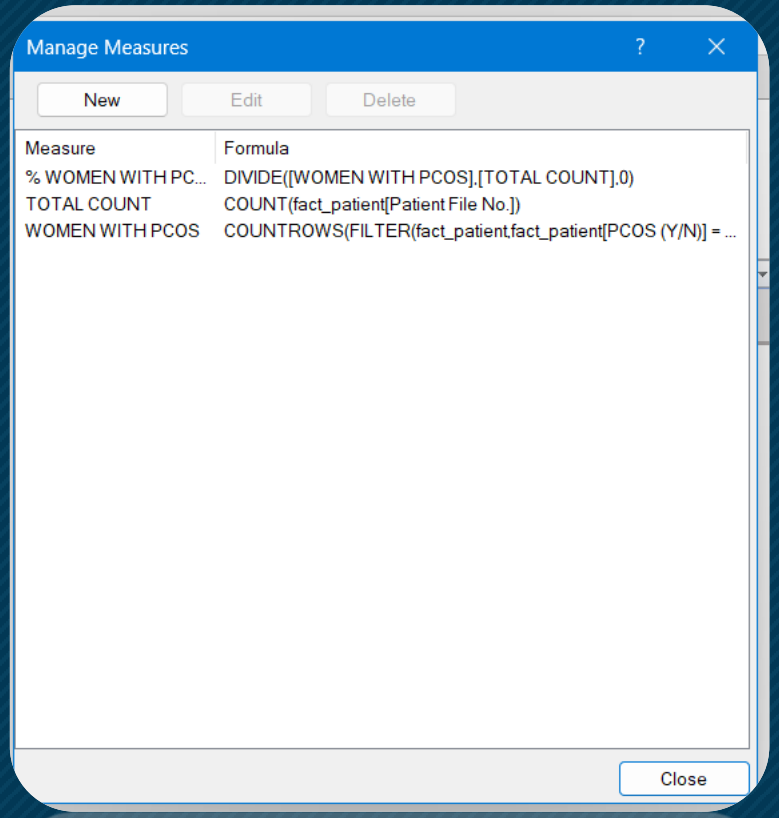
Age range	Women Diagnosed
20-26	33
27-33	59
34-40	29
41-47	4
Grand Total	125

BMI : DIAGNOSED AND % OF TOTAL UNDER RESPECTIVE CATEGORY

BMI categories	Women Diagnosed	Percentage
Underweight	7	26.92%
Normal	48	24.37%
Overweight	52	36.36%
Obese	18	52.94%
Grand Total	125	

Pivot tables were inserted. Relevant columns and rows were pulled into respective pivot tables
Conditional formatting was applied to distinguish between different percentages and to understand the severity better

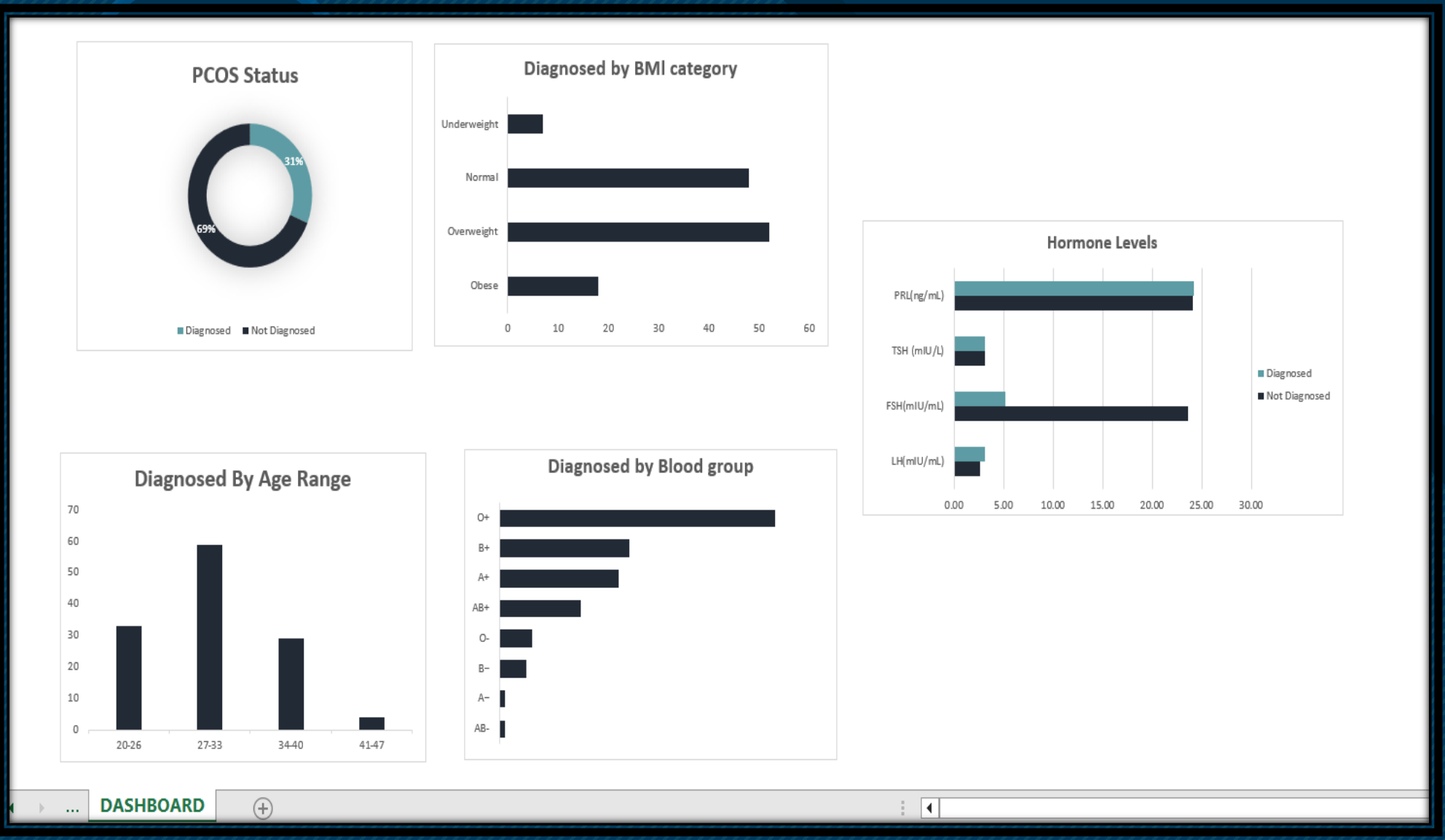
DAX(Data Analysis Expression) MEASURES



Missing columns were identified, hence **DAX measures** were used to create the required formulas.

These measures were then incorporated into the pivot table, enabling the establishment of relationships.

DASHBOARD



This dashboard summarizes PCOS diagnosis trends across key factors:

- Diagnosis status
- BMI, Age, and Blood group distributions
- Hormone level comparisons

These insights highlight key patterns and potential risk indicators.

TO CONCLUDE...

KEY CONCLUSIONS

- **Higher BMI** (specifically women under the overweight category) is linked to increased PCOS cases.
- Most diagnoses occur in the **20–34 age range**.
- **B+ and O+** blood groups show slightly higher occurrence.
- PCOS patients have higher **LH and PRL** hormone levels in comparison to non-diagnosed women.
- **DAX measures** and dashboards helped highlight these patterns effectively.

Thank You

DATASET SOURCE

https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos?select=PCOS_data_without_infertility.xlsx