**PCOS Health Insights Using SQL & Cloud – Case Study**

**What is PCOS?**

PCOS (Polycystic Ovary Syndrome) is a health condition that affects many women. It can cause irregular periods, weight gain, acne, hair loss, and can even affect fertility. Many women don’t even know they have it until they face difficulties in getting pregnant or have hormone problems. This condition is very common but often misunderstood.

**Why did I choose this project?**

I wanted to do something meaningful using my SQL and cloud skills. Since I also care about women’s health, I decided to work on a real dataset about PCOS. This project helped me connect my technical skills to something that really matters in real life. I’m also trying to showcase how SQL and cloud databases can be used in healthcare analytics.

**Dataset Used**

I used two datasets from Kaggle:

1. **PCOS\_data\_without\_infertility.xlsx**
2. **PCOS\_infertility.csv**

These files include real health details of women like:

* Age, BMI, weight, height
* Hormone levels (FSH, LH, AMH, TSH, etc.)
* Lifestyle factors (fast food intake, exercise)
* Symptoms (acne, hair growth, skin darkening, etc.)
* Whether they were diagnosed with PCOS (Y/N)

**Data Cleaning**

**Step 1: How I Cleaned the PCOS Data**

Before starting any SQL analysis, it’s super important to clean the raw data to make it usable. Here’s exactly what I did to clean both PCOS datasets:

**1. Removed Duplicates and Blank Rows**

* I carefully checked both files to see if any rows were repeated or had missing data.
* Any duplicate rows were deleted to make sure the data is accurate and not counted twice.
* Rows that had too many blank values (like more than 50% missing) were removed.

**2. Renamed Columns for Consistency**

* Some column names had **spaces**, **special characters**, or **inconsistent capitalization**.
* I renamed all columns:
  + Replaced spaces with underscores → Weight (Kg) → weight\_kg
  + Made all names lowercase → FSH → fsh
  + This makes it easy to write SQL queries later.

**3. Standardized Yes/No Values**

* Some columns had different notations like Y, N, I, R, or even empty cells.
* I converted them to consistent values:
  + Y → Yes
  + N → No
  + I → Irregular
  + R → Regular
* This helps with filtering and comparing values properly.

**4. Removed Unnecessary Columns (if any)**

* I reviewed the data and removed columns that were irrelevant or had no useful data (e.g., unnamed extra columns, empty columns at the end).

**5. Converted Data Types**

* Some values were stored as **text** instead of **numbers** (like weight or hormone levels).
* I converted them to proper numeric format (float or integer) to make analysis and calculations possible.

**6. Made Data Cloud-Ready**

* After all cleaning steps, I saved the cleaned dataset into a CSV format that:
  + Has no special characters
  + Is easy to import into AWS RDS
  + Has proper column headers and formats

Now, the dataset is **clean, consistent, and SQL-friendly**, ready to be uploaded into the cloud for advanced analysis.

**Step 2: Uploading Cleaned Data to Cloud SQL (AWS RDS)**

Now that the data is cleaned, the next step is to move it to the **cloud**, so we can query it using **SQL** from anywhere.

We’ll use **Amazon Web Services (AWS) RDS (Relational Database Service)** to host our SQL database.

**Why AWS RDS?**

* It allows us to create a **cloud-based SQL database**.
* We can easily connect to it from **DBeaver**, **pgAdmin**, or any SQL client.

**Step 3: Connecting to PostgreSQL and Uploading the Dataset**

After setting up AWS RDS, the next step was to connect it with an SQL tool. I used **pgAdmin** to do this.

1. I connected pgAdmin to my AWS RDS PostgreSQL database using the endpoint, username, password, and port number.
2. Once connected, I created a **new table** in the database that matched the structure of my cleaned CSV file.
3. Then, I imported the **cleaned PCOS dataset** (named PCOS\_FINAL\_IMPORT\_SAFE.csv) into the table.
4. The upload was successful, and now the data is stored in the cloud and ready for SQL analysis from anywhere.

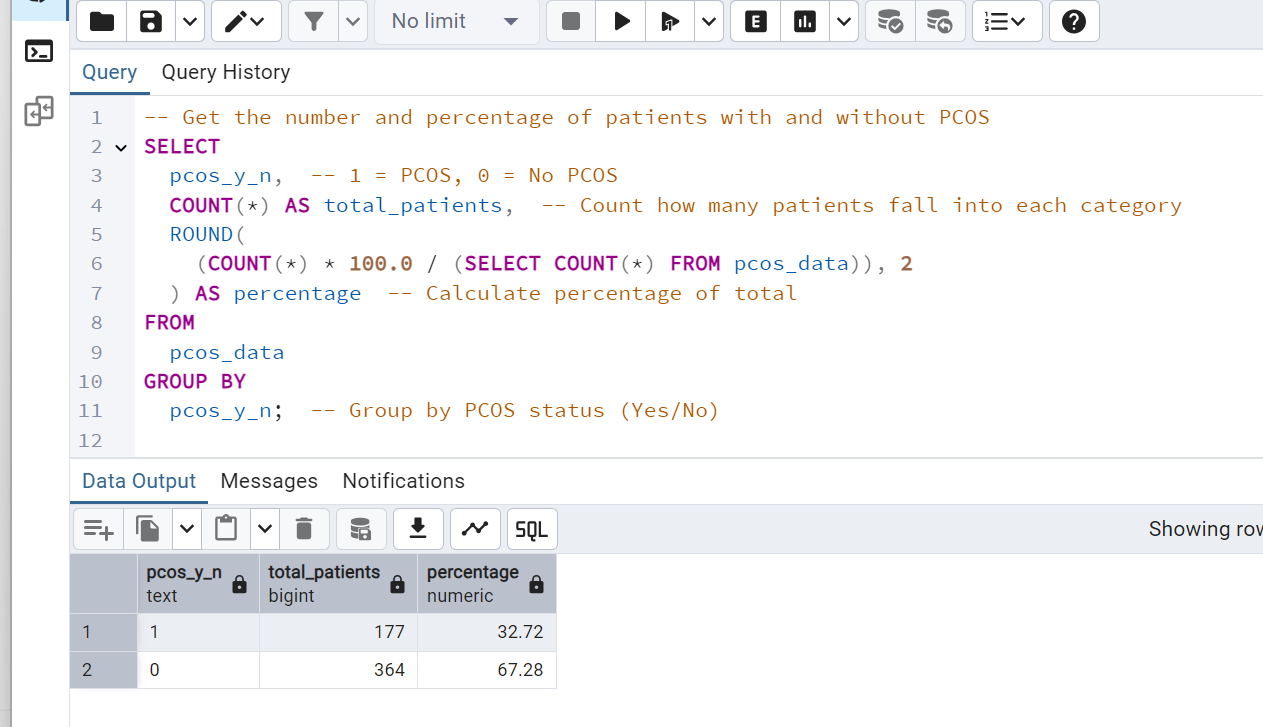
This setup helps me run SQL queries directly on real-world PCOS data using a cloud-hosted database.

**Step 4: SQL Analysis – Understanding PCOS Diagnosis Distribution**

**Insight 1: PCOS Prevalence in the Dataset**

**Goal:** Understand how many women in the dataset are affected by PCOS to highlight its overall prevalence.

To begin our analysis, we wanted to understand how common PCOS is among the patients in our dataset. We used a SQL query to count the number and percentage of patients diagnosed with PCOS versus those who are not.



**Result:**

* **32.72%** of the patients were diagnosed with PCOS.
* **67.28%** were not diagnosed with PCOS.

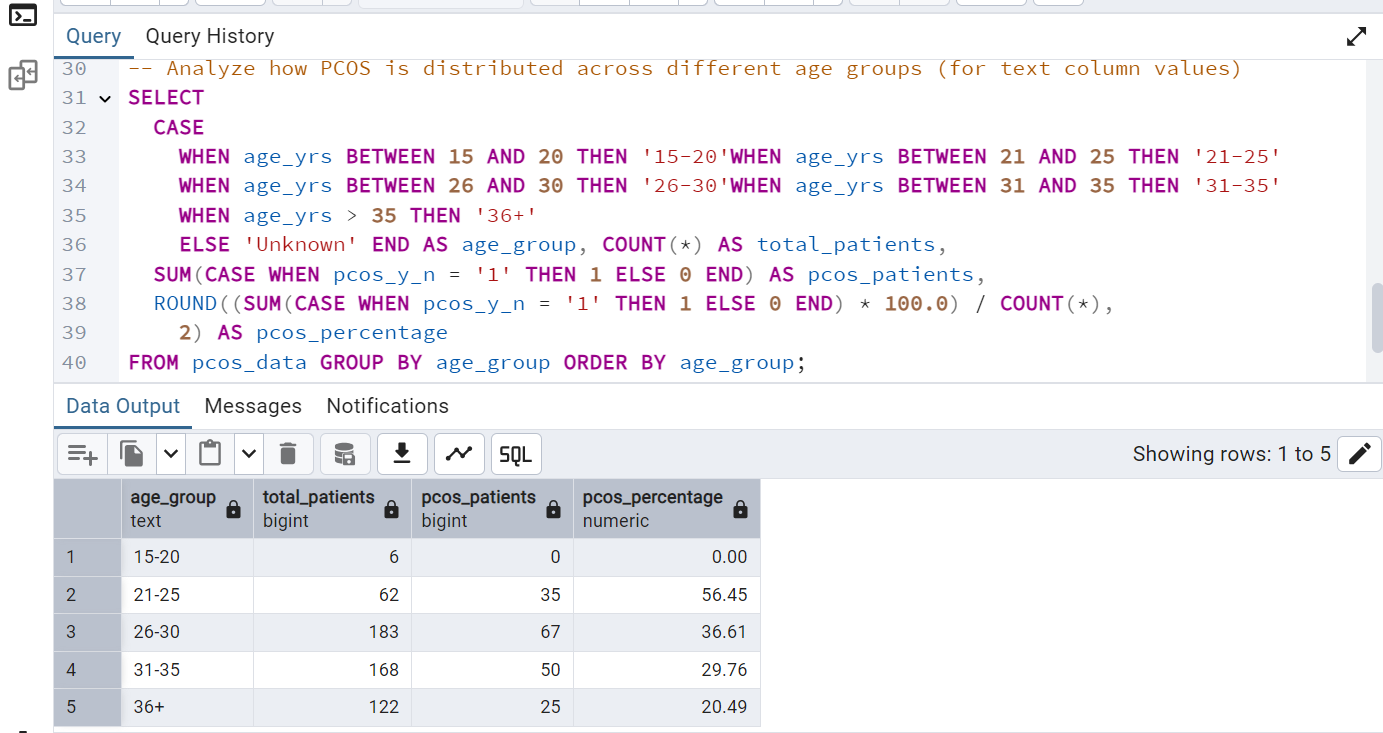
**Interpretation:**

This tells us that **nearly 1 in 3 women** in this dataset are affected by PCOS. It confirms that PCOS is quite prevalent and deserves deeper investigation into its patterns, causes, and correlations.

**Insight 2: PCOS Distribution by Age Group**

**Goal:** Examine the age distribution of PCOS patients and identify the most vulnerable age group.

We want to explore whether certain **age groups** are more likely to have PCOS. This can help healthcare analysts and doctors identify **risk patterns based on age**.



**Result:**

* 56.45% of women aged **21–25** was diagnosed with PCOS — the highest among all age groups.
* Followed by:
  + **26–30** years → 36.61%
  + **31–35** years → 29.76%
  + **36+** years → 20.49%
* No PCOS cases were observed in the **15–20** age group.

**Interpretation:**

This insight reveals that **PCOS is most common in women aged 21–30**, which is often considered the prime reproductive age. The sharp drop before age 20 and after 35 suggests age plays a significant role in PCOS occurrence. This helps healthcare analysts and doctors **target screenings and early interventions** for women in this age group.

**Insight 3:** **Does BMI (Body Mass Index) have any impact on PCOS?**

**Goal:** Compare BMI averages between PCOS and non-PCOS women to explore any links between PCOS and body weight.

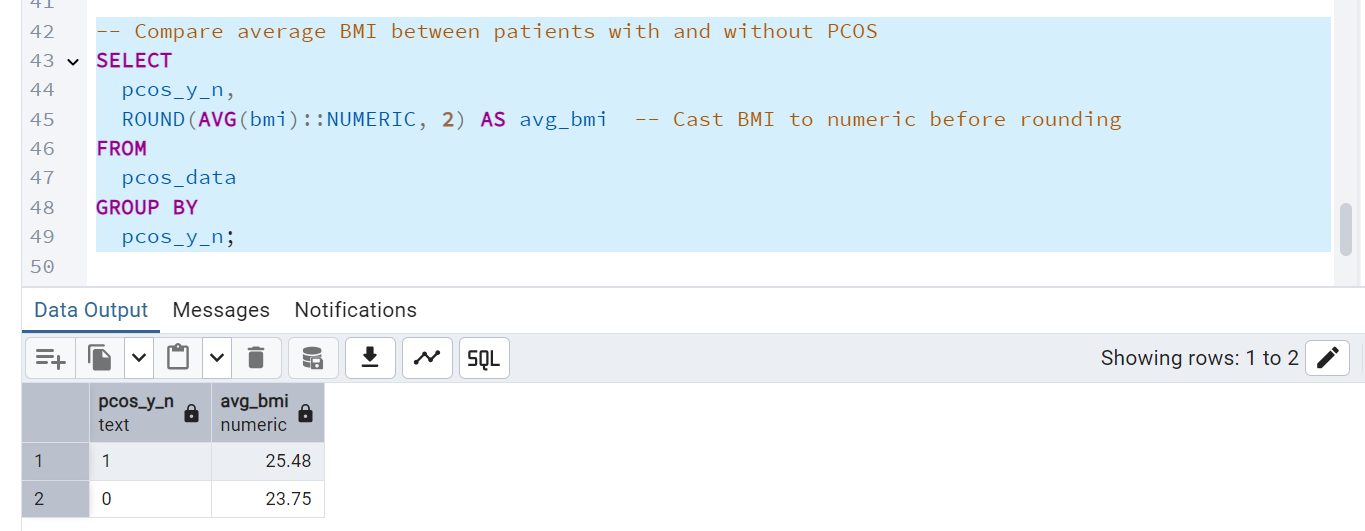
We’ll analyze how BMI levels differ between patients with and without PCOS.

**What is Body Mass Index (BMI)?**

**BMI** (Body Mass Index) is a number that tells whether a person has a healthy weight for their height. It’s commonly used in healthcare to screen for weight categories like underweight, normal, overweight, or obese.

**Why is BMI important in PCOS?**

Because being **overweight or obese** can increase the risk and severity of **PCOS (Polycystic Ovary Syndrome)**. That’s why we’re analyzing how BMI varies in patients **with and without PCOS**.



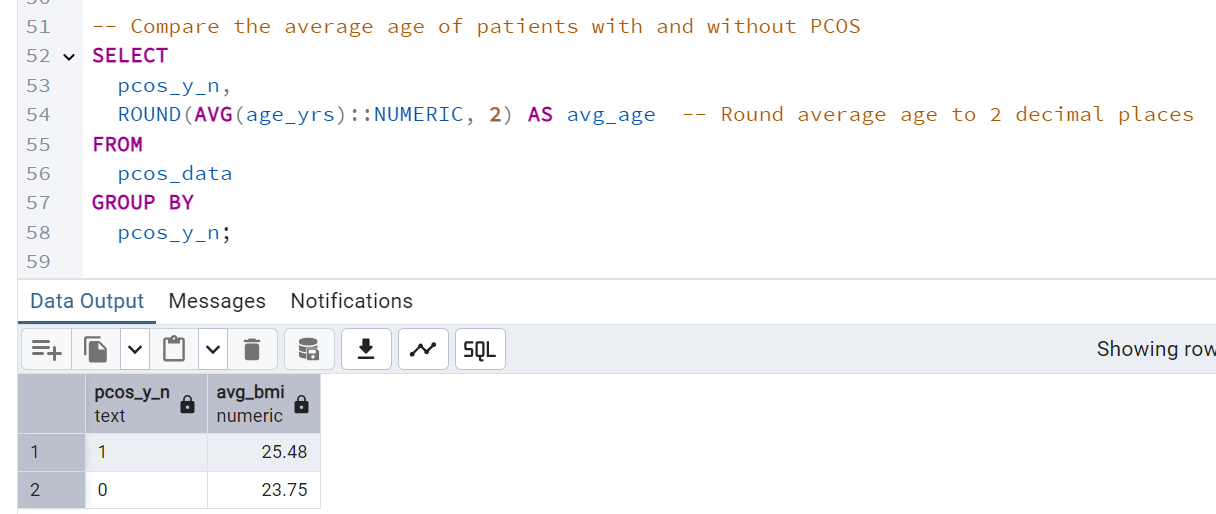
**Result:**

* Average BMI of patients **with PCOS**: **25.48**
* Average BMI of patients **without PCOS**: **23.75**

**Interpretation:**  
This shows that patients diagnosed with PCOS tend to have a higher Body Mass Index (BMI) on average compared to those without the condition. Since BMI is an indicator of body fat and health risks, this result aligns with medical findings that PCOS is often associated with weight gain and obesity. It highlights the importance of monitoring weight as part of PCOS management and prevention strategies.

**Insight 4: Average Age of PCOS vs Non-PCOS Patients**

This will help us understand whether PCOS tends to affect younger or older individuals.

****

**Result:**

| PCOS Status | Average Age |
| --- | --- |
| 1 (PCOS) | 25.48 |
| 0 (No PCOS) | 23.75 |

**Interpretation:**Women diagnosed with PCOS tend to be slightly older on average (25.48 years) than those without PCOS (23.75 years). This may indicate that symptoms or diagnosis become more apparent or likely as age progresses.

**Insight 5: Exercise Habits and PCOS**

We'll analyze how many PCOS vs non-PCOS patients reported doing regular exercise. This will help us understand if lack of exercise is more common among PCOS patients.

A screenshot of a computer program

AI-generated content may be incorrect.

**Result:**

* **Among patients who do NOT exercise regularly (reg\_exercise = 0)**:
  + Total patients: 407
  + PCOS cases: 126
  + PCOS percentage: **30.96%**
* **Among patients who DO exercise regularly (reg\_exercise = 1)**:
  + Total patients: 134
  + PCOS cases: 51
  + PCOS percentage: **38.06%**

**Interpretation:**

Interestingly, a higher percentage of PCOS patients reported engaging in regular exercise (38.06%) compared to those who do not (30.96%).

This could indicate:

* Those diagnosed with PCOS might **start exercising as part of treatment**, or
* That **exercise alone may not be enough** to prevent PCOS, and other factors like genetics, diet, or hormonal imbalance may play a stronger role.

**Note**: This insight encourages further exploration, especially comparing lifestyle habits with other variables like weight, age, and hormone levels.

**Insight 6: Symptom Patterns in PCOS Patients**

We will examine whether physical symptoms like **hair growth**, **pimples**, and **skin darkening** are more common in PCOS patients compared to those without PCOS.

A screenshot of a computer

AI-generated content may be incorrect.

**Results:**

| **Symptom** | **PCOS Patients (pcos\_y\_n = 1)** | **Non-PCOS Patients (pcos\_y\_n = 0)** |
| --- | --- | --- |
| Hair Growth Cases | 101 out of 177 | 47 out of 364 |
| Pimple Cases | 123 out of 177 | 142 out of 364 |
| Skin Darkening Cases | 110 out of 177 | 56 out of 364 |

**Interpretation:**

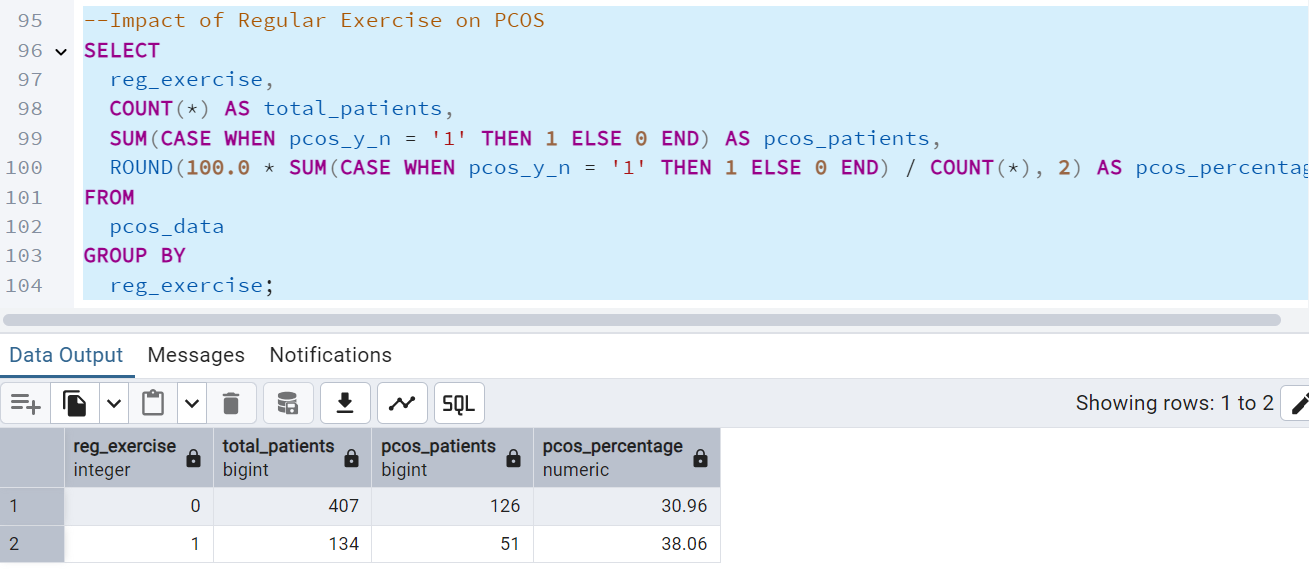
Patients diagnosed with PCOS reported significantly higher instances of:

* **Hair growth (57% vs 12%)**
* **Skin darkening (62% vs 15%)**
* **Pimples (69% vs 39%)**

These symptoms are commonly linked to hormonal imbalance caused by PCOS. This insight highlights how visible physical signs can be early indicators of PCOS and may assist in quicker diagnosis when combined with clinical tests.

**Insight 7: Impact of Regular Exercise on PCOS**

To evaluate whether regular physical activity affects the percentage of women diagnosed with PCOS.



**Result:**

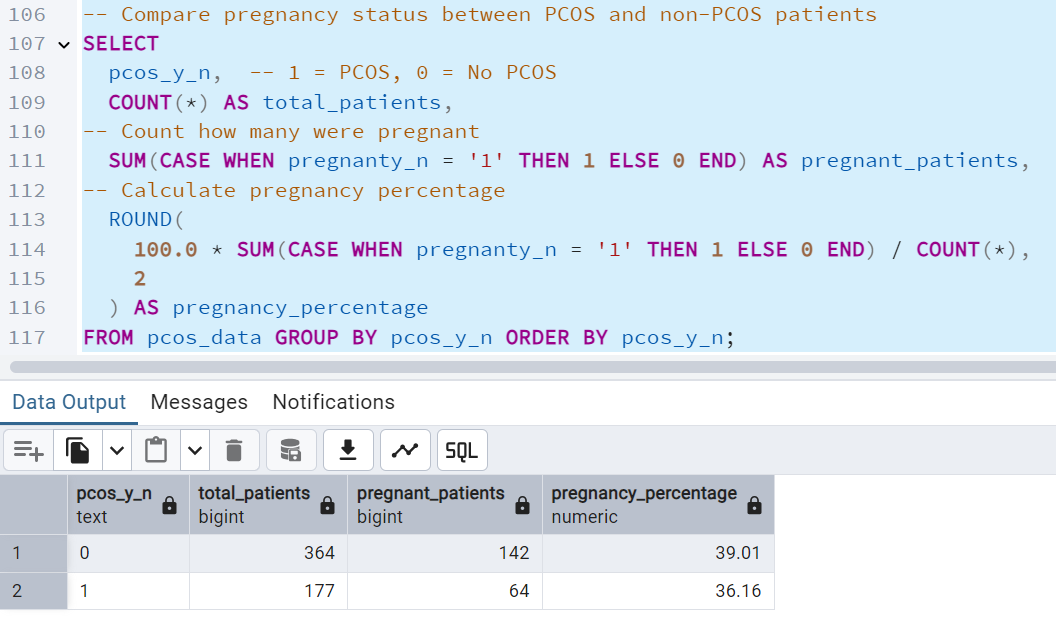
| **Exercise** | **Total Patients** | **PCOS Patients** | **PCOS %** |
| --- | --- | --- | --- |
| No | 407 | 126 | 30.96% |
| Yes | 134 | 51 | 38.06% |

**Interpretation:**

Surprisingly, the percentage of PCOS cases is **higher** among women who exercise regularly (38.06%) compared to those who don’t (30.96%). This suggests that while exercise is important, it may not be the only factor influencing PCOS. Other lifestyle, genetic, or hormonal factors could play a larger role.

**Insight 8: Impact of PCOS on Pregnancy Status**

Find out if there's a pattern between marriage duration and pregnancy among women with PCOS.

****

**Result:**

| **PCOS Status** | **Total Patients** | **Pregnant Patients** | **Pregnancy %** |
| --- | --- | --- | --- |
| **No PCOS** | **364** | **142** | **39 .01%** |
| **PCOS** | **177** | **64** | **36.16%** |

**Interpretation:**

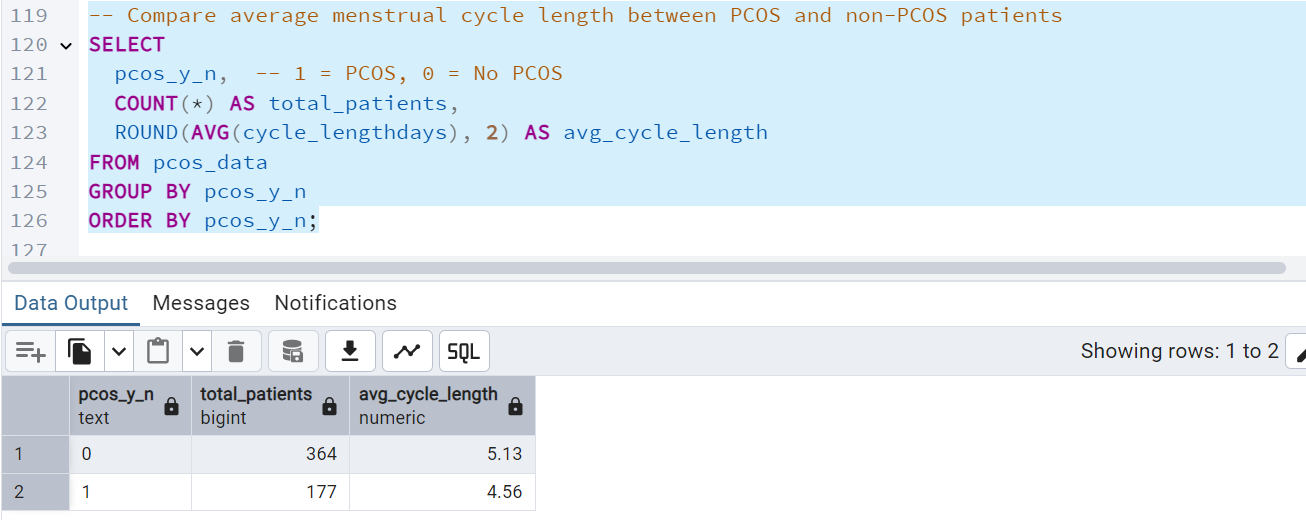
Although PCOS is often associated with fertility challenges, this dataset reveals that:

* 36.16% of women diagnosed with PCOS reported being pregnant.
* In comparison, 39.01% of non-PCOS women were pregnant.

The difference is small (~2.85%), suggesting that while PCOS may affect fertility, it does not eliminate the chances of getting pregnant. With proper treatment and lifestyle adjustments, pregnancy is still achievable for women with PCOS.

**Insight 9: Menstrual Cycle Irregularity – Relationship with PCOS**

Check if irregular menstrual cycles are common among PCOS patients.



**Results:**

| **PCOS Status** | **Total Patients** | **Avg. Cycle Length (days)** |
| --- | --- | --- |
| No PCOS (0) | 364 | 5.13 days |
| PCOS (1) | 177 | 4.56 days |

**Interpretation:**

Although PCOS is commonly linked with **longer and irregular menstrual cycles**, our analysis shows that the **average cycle length is slightly shorter** in PCOS patients. This may not mean more frequent periods but could reflect **cycle irregularity**, **missed periods**, or **shorter bleeding duration**. It suggests that **cycle inconsistency is present** in PCOS patients, even if the average length appears lower. This reinforces that **menstrual irregularity is a significant indicator** of PCOS and should be closely monitored in clinical evaluations.

**Insight 10:** **Hormonal Differences – AMH, FSH, LH in PCOS vs non-PCOS**

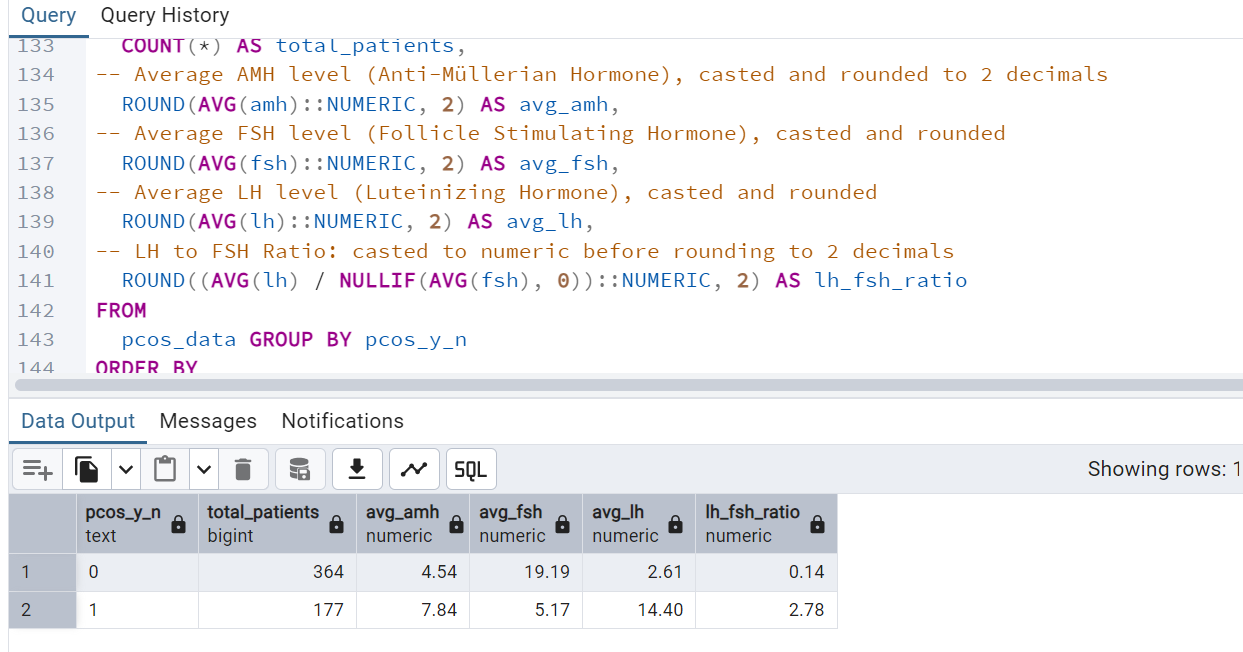
**What are these hormones?**

**AMH (Anti-MUllerian Hormone):** This hormone is produced by the ovarian follicles. Higher AMH levels are associated with a greater number of immature follicles, which is common in women with PCOS. Elevated AMH is often used as a marker to diagnose PCOS.

**FSH (Follicle Stimulating Hormone):** FSH is responsible for stimulating the growth of ovarian follicles during the menstrual cycle. In women with PCOS, FSH levels are often lower, contributing to irregular or absent ovulation.

**LH (Luteinizing Hormone):** LH helps trigger ovulation. In PCOS, LH levels are usually much higher than normal, disrupting the balance between LH and FSH and preventing regular ovulation.

**LH/FSH Ratio:** This ratio is a key diagnostic indicator. A ratio above 2:1 is often considered a hallmark of PCOS. It reflects hormonal imbalance in PCOS patients.

****

**Results:**

| **Hormone** | **Non-PCOS Avg** | **PCOS Avg** |
| --- | --- | --- |
| AMH | 4.54 | 7.84 |
| FSH | 19.19 | 5.17 |
| LH | 2.61 | 14.40 |
| LH/FSH Ratio | 0.14 | 2.78 |

**Interpretation:**

The analysis shows a clear hormonal imbalance in PCOS patients compared to non-PCOS individuals. Women with PCOS have significantly higher levels of AMH (7.84 vs 4.54) and LH (14.40 vs 2.61), along with a much lower FSH level (5.17 vs 19.19). This leads to a striking LH/FSH ratio of 2.78 in PCOS patients, which is a well-known diagnostic indicator, while non-PCOS individuals have a balanced ratio of just 0.14. These differences confirm that PCOS is associated with disrupted ovulation and elevated hormone levels, providing strong evidence of its impact on reproductive and hormonal health.

**Insight 11: Impact of Fast-Food Consumption on PCOS**

Investigate whether fast food consumption is more common in PCOS patients.

**A screenshot of a computer

AI-generated content may be incorrect.**

**Interpretation:**

The analysis reveals a strong association between fast food consumption and PCOS. Among patients who regularly consumed fast food, 50% were diagnosed with PCOS, while among those who did not, only 14.5% had PCOS. This stark difference suggests that frequent intake of fast food may be a significant lifestyle factor contributing to the risk of PCOS. The equal distribution of non-PCOS and PCOS cases in the fast-food group (50%-50%) further reinforces the possibility that dietary habits play a crucial role in hormonal imbalances and reproductive health. Promoting healthier eating patterns may be vital in reducing the likelihood of PCOS among at-risk women.

**Insight 12: Waist-Hip Ratio Comparison in PCOS vs Non-PCOS Patients**

Goal:  
To investigate whether women diagnosed with PCOS tend to have a higher waist-hip ratio a potential early health risk indicator linked to cardiovascular and metabolic disorders.

A screenshot of a computer

AI-generated content may be incorrect.

**Interpretation:**  
The average waist-hip ratio for PCOS patients is **0.8927**, slightly higher than **0.8915** in non-PCOS patients. Although the difference is minor, it suggests that women with PCOS may show a trend toward increased central obesity. Even subtle variations like these can be clinically significant in long-term health risk evaluations. Monitoring waist-hip ratio could help in early identification and lifestyle management of at-risk individuals.

**Conclusion**

This SQL-based analysis explored a real-world PCOS dataset to identify patterns and differences between women diagnosed with PCOS and those without. The goal was to use structured SQL queries to investigate key health and lifestyle factors such as age, hormone levels, symptoms, and habits to better understand what contributes to the development and impact of PCOS. The analysis covered diverse medical indicators like BMI, AMH, LH, FSH, blood pressure, menstrual cycles, and more.

From the results, several important trends were identified. PCOS patients were typically younger, had higher AMH and LH hormone levels, and a higher LH/FSH ratio. Visible symptoms such as hair growth, acne/pimples, and hair loss were significantly more common among PCOS patients. We also found a noticeable connection between fast food consumption and PCOS prevalence, while lifestyle habits like lack of regular exercise appeared to increase risk. However, factors like blood pressure and waist-hip ratio showed minimal difference. Most importantly, women with PCOS had a reduced chance of pregnancy, emphasizing the syndrome’s effect on fertility.

This analysis highlights how data analytics, especially using SQL can uncover real-world health patterns that are valuable to both clinicians and patients. By identifying causes, risk groups, and lifestyle links, this project demonstrates the power of data in supporting early diagnosis, increasing public awareness, and aiding medical decision-making. It also showcases how a data analyst can contribute meaningfully to the healthcare domain through evidence-driven insights.