



Impact of Lifestyle and Physiological Factors on PCOS: A Statistical Analysis

INSTRUCTOR: Gopikrishnan Chandrasekharan Ph.D

TEAM MEMBERS:

Archita Singamsetty

Jamesetta Quiqui

Treyden Stansfield

Veera Venkata Satyavathi Surapureddy

INTRODUCTION



DATASET: PCOS Dataset - Kaggle

Problem Statement: Polycystic Ovary Syndrome (PCOS) is an endocrine disorder that affects women in the reproductive age group and is characterized by ovarian follicle dysfunction, elevated androgen levels, and irregular menstrual cycles. Due to its association with a number of metabolic, hormonal and reproductive complications, it is important to determine possible predictors and risk factors for the condition to enable early diagnosis and management. Some of the factors include Body Mass Index (BMI), thyroid function through **FSH/LH** levels, The blood systematic pressure review and and diet meta-analysis and by exercise habits.

Addressing the Problem Statement Using Statistical Analysis: Our study uses statistical analysis to explore these predictors and assess how they are linked to PCOS (Polycystic Ovary Syndrome). By utilizing techniques, like **Exploratory Data Analysis (EDA)**, **Chi-Square tests**, and **Logistic Regression analysis** tools we seek to reveal connections between factors and lifestyle choices with the occurrence of PCOS. This research offers evidence based findings that can improve our knowledge of PCOS risk factors leading to strategies, for dealing with and preventing the condition.

Hypothesis



- Women with unhealthy lifestyles, such as high fast-food consumption and low levels of exercise, combined with physiological imbalances (e.g., high BMI, high blood pressure, and abnormal hormone levels), are at a significantly higher risk of developing PCOS.
- These factors not only affect the likelihood of developing PCOS but also influence its severity.

Testing Approach:

We have used **Chi-Square Test** to analyse the relationships between categorical variables such as exercise, diet and PCOS to determine the significance of these relationships. **Logistic Regression** was used to understand how various physiological and lifestyle parameters are interrelated to the risk of PCOS. **Descriptive Statistics and EDA** were also used to calculate general characteristics of the data and to explore the distribution of the data to support the current analysis.

Dataset Description



Our analysis is performed on a **PCOS Dataset - Kaggle** that is available to the public and includes **2,000 data points** and **44 variables**. The data set includes the clinical, demographic, and lifestyle data that may have relationship with the incidence of PCOS. Therefore, this dataset is a great opportunity to gain deeper insights into the correlations between the physiological and behavioral factors and PCOS.

Variables can be classified into different categories depending on their characteristics:

Continuous Variables:

- **BMI (Body Mass Index):** An index of body fat which is derived on the basis of weight and height, which is a continuous variable.
- **FSH/LH (Follicle-Stimulating Hormone over Luteinizing Hormone):** A ratio hormonal marker of PCOS
- **RBS (Resting Blood Sugar):** Average blood sugar level
- **PRG (Progesterone):** Opposite of Androgens and is a female hormone
- **Systolic and Diastolic Blood Pressure:** Proxies of cardiovascular disease, measured in mm Hg and therefore continuous variables.

Categorical Variables:

- **PCOS (Yes/No):** It is the variable that shows the presence or absence of PCOS and it is a binary variable.
- **Weight Gain (Yes/No):** Self-assessment of weight gain as a binary variable.
- **Regular Exercise (Yes/No):** Shows whether people engage in exercise regimen (s) or not (binary).
- **Fast Food Consumption (Yes/No):** It measures the consumption of fast food which can be in few times a week or even daily (binary)

Dataset Description



Variable Distribution

For continuous variables:

- BMI varies from 11.94 to 40.45 with a mean of 24.28.
- TSH has values between 0.04 and 65.00 with the mean of 3.00.
- Blood pressure levels are within the normal range as seen in the general population although with a wide variation and the systolic pressure was 114.

For categorical variables:

- The dataset is balanced with about one third of the participants being diagnosed with PCOS. The dataset is balanced with approximately 30% of individuals diagnosed with PCOS.
- About 38 % of women experienced weight gain, 25 % engage in exercise, and 52 % take fast food.

Dataset Description



Visual Summaries

To understand the data distribution:

- **Histograms** were made for quantitative variables such as BMI, TSH, Blood Pressure and AMH to show that the data had a normal or slightly skewed distribution.
- **Bar Charts** were developed for categorical data such as PCOS, weight gain, exercise and fast food consumption to determine the frequency of each category.
- **Correlation Heatmap:** The correlation between all the continuous variables were calculated and a correlation matrix was created and then represented as a heatmap to identify the strength of the linear relationship between the variables which included BMI, TSH, systolic blood pressure, and diastolic blood pressure.

Sampling Techniques:

The present dataset does not state the method of sampling used for data collection. However, it seems to contain a number of people who may develop PCOS, which makes the dataset relevant for identifying variables that are associated with the condition. In order to make the analysis consistent, the missing values in the continuous variables were filled with the **mean**, which allowed us to preserve the data integrity and at the same time make the analysis easier.

Statistical Methods



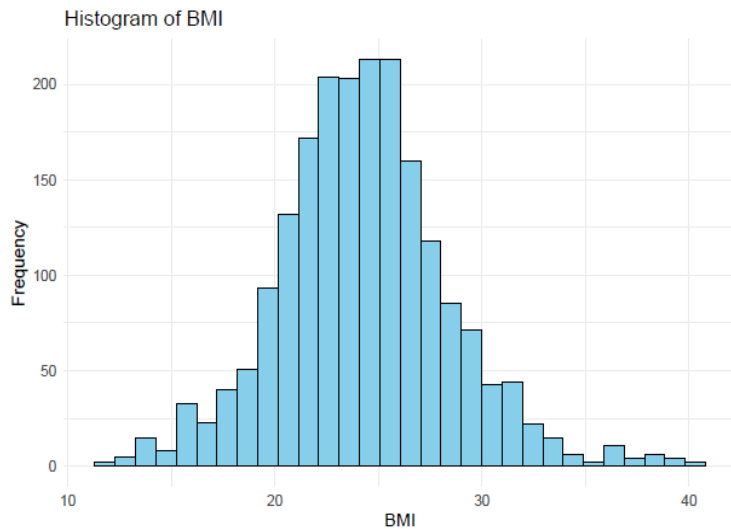
Exploratory Data Analysis (EDA):

- In our project, Exploratory Data Analysis (EDA) was conducted to uncover patterns and relationships within the data. This included summarizing the data through measures like means.. Visualization techniques, such as histograms and bar charts, were used to understand the distribution of variables, identify trends, and detect potential outliers.
- In order to make our regression model more robust and to deal with multicollinearity among the predictors, we have conducted **Variance Inflation Factor (VIF) analysis**. **Multicollinearity** is a phenomenon whereby the independent variables are correlated and this can cause inflation of standard errors thereby making it difficult to make accurate estimates of the effects of the predictors. Generally, **VIF values that exceed 5** are deemed as having severe multicollinearity issues which implies that there is high level of correlation between the variables.

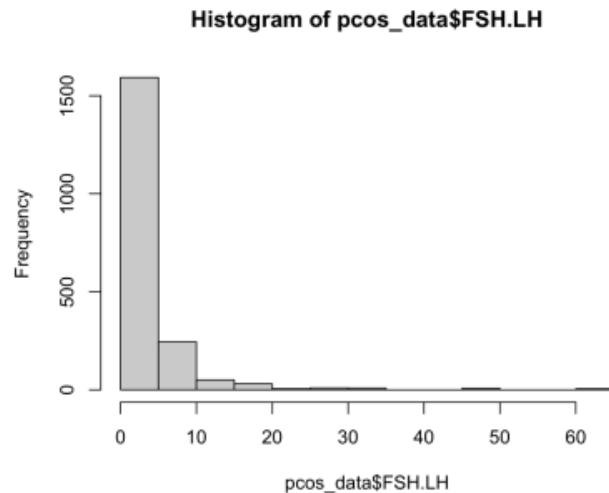
Variable	VIF value
FSH.LH	1.006357
BP._Systolic..mmHg.	1.041913
BP._Diastolic..mmHg.	1.042729
RBS.mg.dl.	1.006886
PRG.ng.mL.	1.008715

Visualizations

Visualizing the spread of BMI among participants.

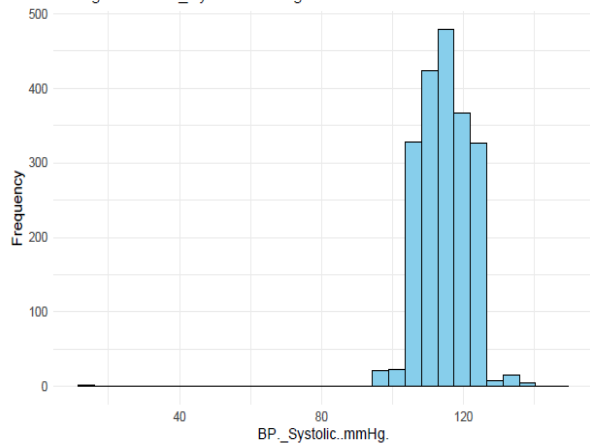


Highlighting the variation in FSH/LH levels



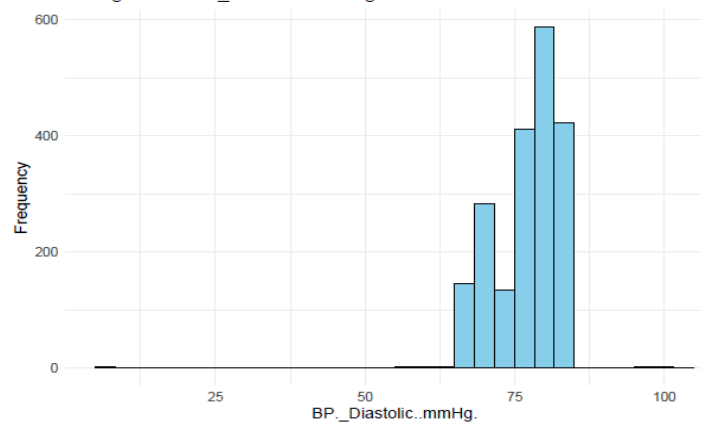
- Examining systolic blood pressure among participants

Histogram of BP_Systolic..mmHg.

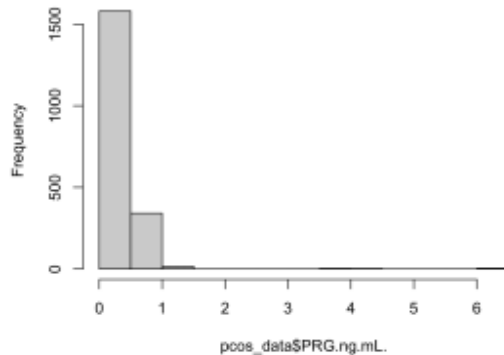


- Analyzing diastolic blood pressure variations.

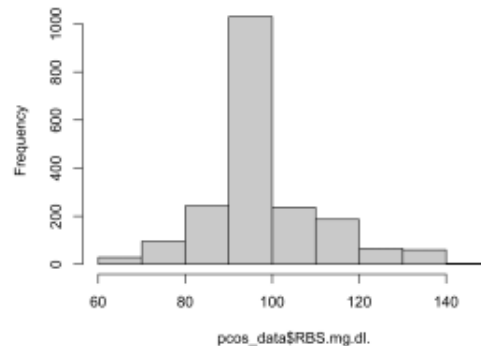
Histogram of BP_Diastolic..mmHg.



Histogram of pcos_data\$PRG.ng.mL.



Histogram of pcos_data\$RBS.mg.dl.

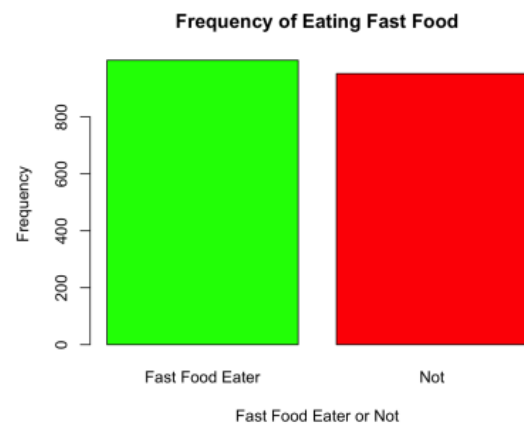
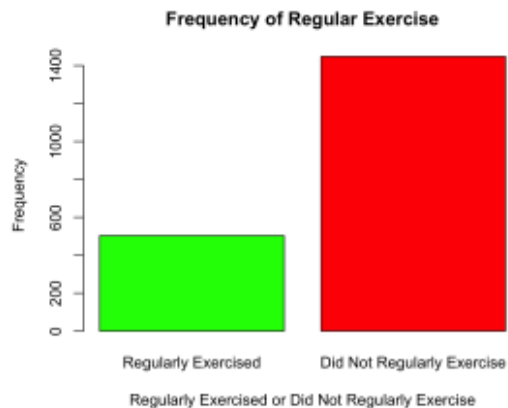
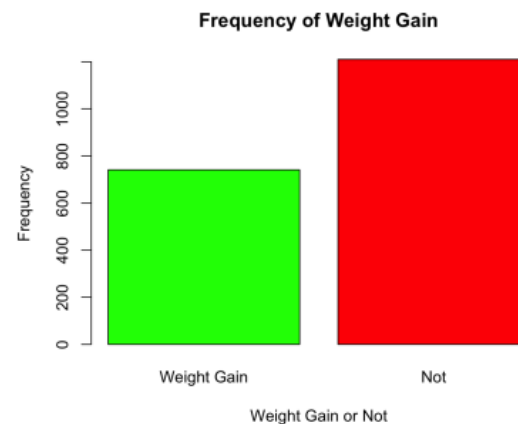
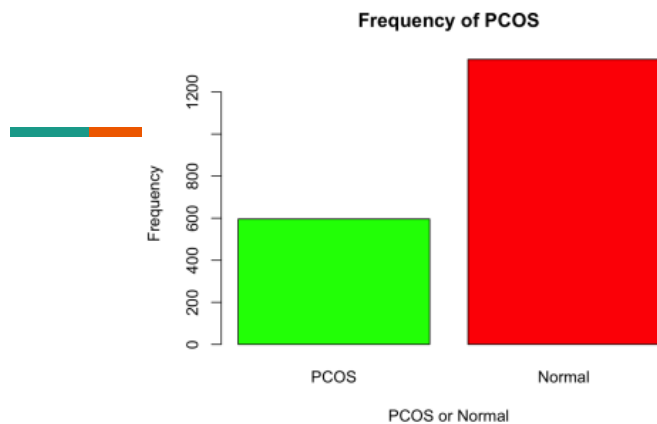


Interpretation of the histograms



The histograms presented summarize the distribution of continuous variables used in our analysis.

- **BMI:** Displays a near-normal distribution with a concentration around 20–30.
- **TSH (Thyroid Stimulating Hormone):** Skewed towards lower values, with most observations below 5 mIU/L.
- **Systolic and Diastolic Blood Pressure:** Both show normal-like distributions, with systolic values centered around 110-120 mmHg and diastolic values around 70-80 mmHg.
- **AMH (Anti-Müllerian Hormone):** Positively skewed, with a majority of values below 10 ng/mL.



Interpretation of the bar charts

The bar charts provide insights into the distribution of categorical variables in the dataset:

- **PCOS (Y/N):** The majority of the individuals do not have PCOS, with a smaller proportion having PCOS.
- **Weight Gain (Y/N):** A significant number of individuals did not experience weight gain, although a notable proportion did.
- **Regular Exercise (Y/N):** Most individuals reported not engaging in regular exercise, while a smaller group did.
- **Fast Food (Y/N):** The distribution of fast food consumption is almost evenly split between individuals who consume it and those who do not.

Statistical Methods



Correlation Analysis:

- A correlation heatmap was constructed in order to assess the presence of multicollinearity among continuous explanatory variables.
- This was further supported by Low Variance Inflation Factors ($VIF < 1.05$) which showed that there was no severe multicollinearity which makes the model reliable.

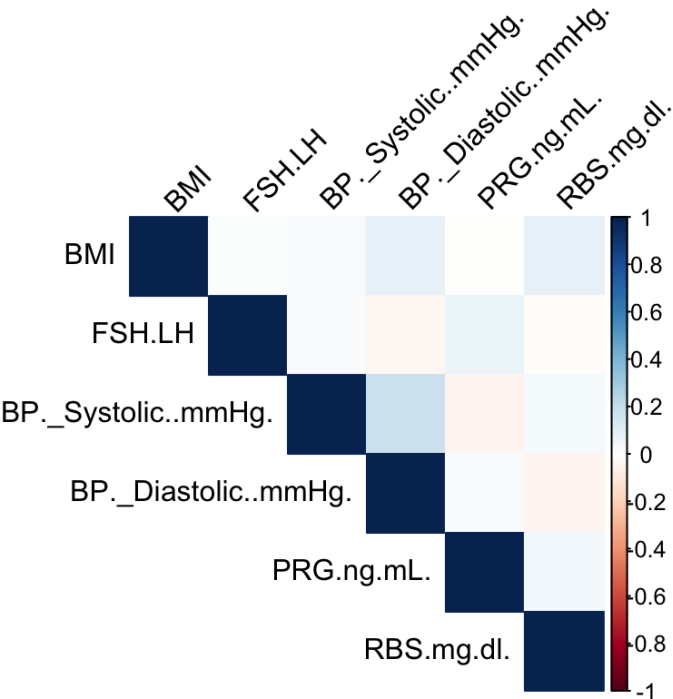
Limitations:

- The statistical analysis is very much influenced by the data that are used in it. It was also important to deal with missing values and outliers to get precise outcomes.
- One of the major limitations of the study is that logistic regression assumes linearity which may not capture the non-linear interactions between the predictors.
- Chi-Square tests can only be performed on categorical data which may not always be the best way of representing the relationship between variables as they may not capture all the possible categories.


Correlation Heatmap for Continuous Variables



Correlation Matrix of Predictors for Predicting BMI



Statistical Methods Logistical regression Anaylsis



Variable	Coefficient	P-value
BMI	0.01788	0.65212
Weight.gain.Y.N. Yes	1.55017	< 2e-16
Fast.food..Y.N. Yes	1.32378	< 2e-16
Reg.Exercise.Y.N. Yes	0.67961	1.4e-07
FSH.LH	-0.007825	0.45123
BP._Systolic..mmHg.	-0.017433	0.04662
BP._Diastolic..mmHg.	-0.012217	0.29691
RBS.mg.dl.	-0.002613	0.54611
PRG.ng.mL.	-0.823928	0.00471

Statistical Methods



Chi-Square Test: To understand the correlation between categorical predictors, chi-square tests were conducted to examine the relationship between lifestyle factors such as weight gain, fast food consumption and regular exercise and the occurrence of PCOS. Bonferroni adjustment (p-critical value = 0.0167) helped in maintaining a very conservative level of significance thus reducing the chances of Type I error. The results showed that there was a very high level of relationship between PCOS and weight gain ($p < 0.0167$), fast food ($p < 0.0167$) and exercise ($p = 0.002$).

Chi Square test with a bonferroni adjusted
p-critical value of 0.0167 (0.05/3) :

Test	Test Stat	P-value
Weight Gain vs PCOS	370.17	$< 2.2e-16$
Fast Food vs PCOS	293.96	$< 2.2e-16$
Reg exercise vs PCOS	8.97	0.002

Discussion and Conclusion



- Of the significant factors, three were lifestyle related (Fast food Consumption, Weight Gain, Blood Pressure and Regular Exercise) and one was hormonal (Progesterone levels).
- Weight gain may have caused an increase risk of PCOS as women who have recently gained weight may work out more to counter this.
- It seems lifestyle factors related to **Weight gain are a better predictor of PCOS status than Hormonal factors.**