 Endometriosis Diagnosis Analysis Using Python

Project Overview

Endometriosis is a chronic medical condition that affects approximately 10% of women of reproductive age. In this project, I analyze a dataset of 10,000 patients to identify key factors influencing diagnosis. Using Python and data visualization, I explore relationships between features like chronic pain, hormone abnormalities, and menstrual irregularities.

Objective

The goal of this project was to analyze medical data related to endometriosis and uncover significant factors contributing to diagnosis. By leveraging Python (Pandas, Seaborn, Matplotlib), I conducted exploratory data analysis (EDA) and correlation analysis to highlight the most relevant predictors of the condition.

```
#import necessary libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

Data Loading

```
df = pd.read_csv("C:/Users/KIIT/Desktop/Projects/Endometriosis
dataset/structured_endometriosis_data.csv")

df
```

```
      Age  Menstrual_Irregularity  Chronic_Pain_Level  \
0      24                       1            8.361531
1      37                       1            4.995508
2      46                       1            3.363996
3      32                       1            5.246037
4      28                       1            3.898932
...   ...                     ...                 ...
9995   37                       1            3.153169
9996   44                       0            4.044800
9997   39                       0            5.096384
9998   47                       1            7.598862
9999   38                       0            7.822210

      Hormone_Level_Abnormality  Infertility        BMI  Diagnosis
0                             0            0  19.451314          0
1                             0            0  22.388436          0
2                             1            0  21.320443          0
3                             0            0  20.177715          1
4                             1            0  23.538103          1
...                         ...          ...        ...        ...
9995                          1            0  18.318849          0
```

```
9996                            1            1  24.732344          0
9997                            1            1  34.204883          1
9998                            1            1  30.374964          1
9999                            0            0  26.385575          0

[10000 rows x 7 columns]
```

Understanding the Dataset

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 7 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   Age                         10000 non-null  int64  
 1   Menstrual_Irregularity      10000 non-null  int64  
 2   Chronic_Pain_Level          10000 non-null  float64
 3   Hormone_Level_Abnormality   10000 non-null  int64  
 4   Infertility                 10000 non-null  int64  
 5   BMI                         10000 non-null  float64
 6   Diagnosis                   10000 non-null  int64  
dtypes: float64(2), int64(5)
memory usage: 547.0 KB

#no. of rows , columns
df.shape

(10000, 7)

#first 10 rows
df.head(10)

   Age  Menstrual_Irregularity  Chronic_Pain_Level
Hormone_Level_Abnormality  \
0   24                       1            8.361531
0
1   37                       1            4.995508
0
2   46                       1            3.363996
1
3   32                       1            5.246037
0
4   28                       1            3.898932
1
5   25                       0            3.766540
0
6   46                       1            5.866368
0
```

```
7   38                          1             1.792541
0
8   24                          0             2.112142
1
9   43                          1             5.038582
1
```

```
   Infertility        BMI  Diagnosis
0            0  19.451314          0
1            0  22.388436          0
2            0  21.320443          0
3            0  20.177715          1
4            0  23.538103          1
5            0  24.453548          0
6            0  24.003088          1
7            0  22.590957          0
8            0  24.644436          0
9            0  23.968657          1
```

```
#last 10 rows
df.tail(10)
```

```
      Age  Menstrual_Irregularity  Chronic_Pain_Level  \
9990   49                       1            4.150713
9991   21                       1            5.824441
9992   39                       0            4.763224
9993   35                       1            8.610709
9994   29                       1            6.172818
9995   37                       1            3.153169
9996   44                       0            4.044800
9997   39                       0            5.096384
9998   47                       1            7.598862
9999   38                       0            7.822210
```

```
      Hormone_Level_Abnormality  Infertility        BMI  Diagnosis
9990                          1            0  27.286842          1
9991                          1            0  20.271897          1
9992                          1            1  31.852896          0
9993                          1            0  18.793566          0
9994                          1            1  22.086119          1
9995                          1            0  18.318849          0
9996                          1            1  24.732344          0
9997                          1            1  34.204883          1
9998                          1            1  30.374964          1
9999                          0            0  26.385575          0
```

```
#name of the columns
df.columns
```

```
Index(['Age', 'Menstrual_Irregularity', 'Chronic_Pain_Level',
       'Hormone_Level_Abnormality', 'Infertility', 'BMI',
'Diagnosis'],
      dtype='object')
```

```
#to know the datatypes of the columns
df.dtypes
```

```
Age                             int64
Menstrual_Irregularity          int64
Chronic_Pain_Level            float64
Hormone_Level_Abnormality       int64
Infertility                     int64
BMI                           float64
Diagnosis                       int64
dtype: object
```

```
#the unique values of all the columns
df.nunique()
```

```
Age                              32
Menstrual_Irregularity            2
Chronic_Pain_Level             9875
Hormone_Level_Abnormality         2
Infertility                       2
BMI                            9776
Diagnosis                         2
dtype: int64
```

Data Cleaning

```
#to check the null values
df.isnull().sum()
```

```
Age                             0
Menstrual_Irregularity          0
Chronic_Pain_Level              0
Hormone_Level_Abnormality       0
Infertility                     0
BMI                             0
Diagnosis                       0
dtype: int64
```

```
#checking for the duplicates
df.duplicated().sum()
```

```
0
```

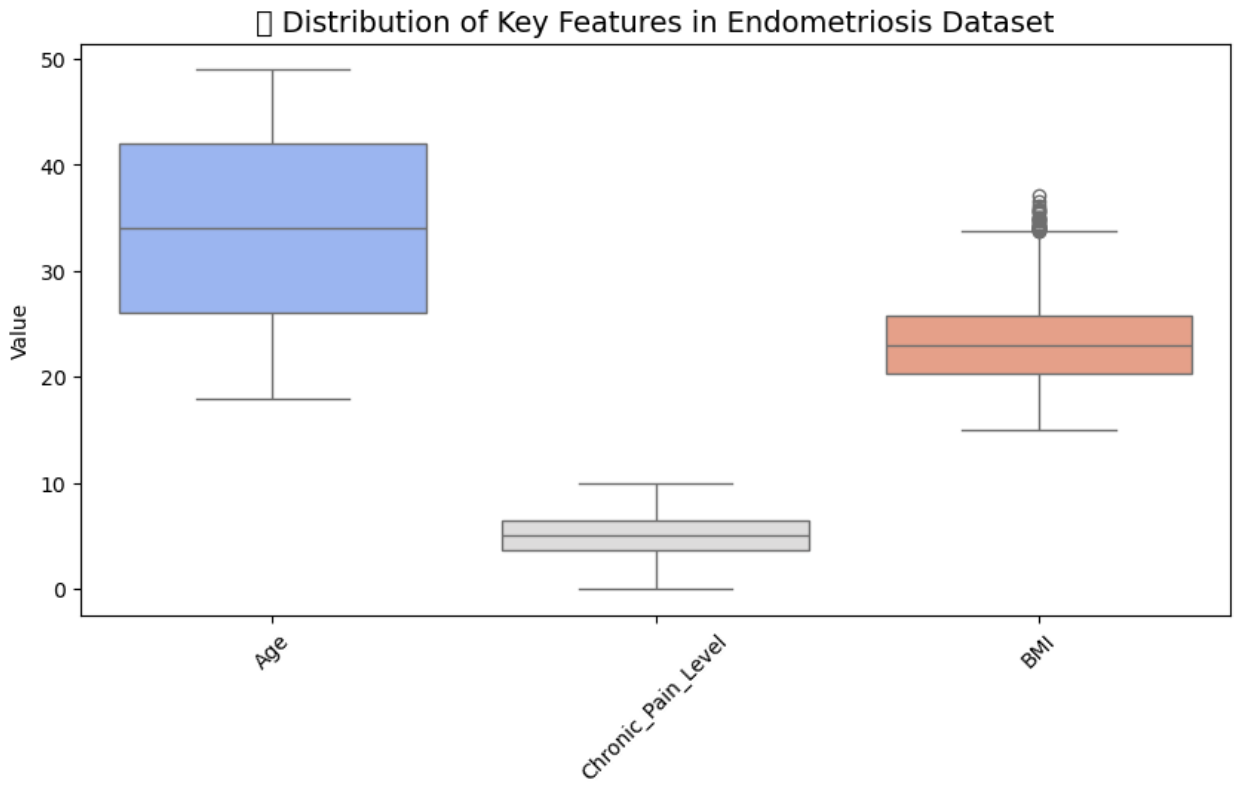
```
#Description of the data
df.describe()
```

```
                Age  Menstrual_Irregularity  Chronic_Pain_Level  \
count  10000.000000            10000.000000        10000.000000
mean      33.692300                0.697500            5.030619
std        9.205308                0.459364            1.983955
min       18.000000                0.000000            0.000000
25%       26.000000                0.000000            3.671697
50%       34.000000                1.000000            5.035825
75%       42.000000                1.000000            6.396854
max       49.000000                1.000000           10.000000

       Hormone_Level_Abnormality   Infertility           BMI
Diagnosis
count               10000.000000  10000.000000  10000.000000
10000.000000
mean                    0.591100      0.298300     23.052865
0.407900
std                     0.491655      0.457535      3.891615
0.491469
min                     0.000000      0.000000     15.000000
0.000000
25%                     0.000000      0.000000     20.329327
0.000000
50%                     1.000000      0.000000     23.036315
0.000000
75%                     1.000000      1.000000     25.712923
1.000000
max                     1.000000      1.000000     37.146127
1.000000
```

Exploratory Data Analysis(EDA)

```
plt.figure(figsize=(10, 5))
sns.boxplot(data=df[['Age', 'Chronic_Pain_Level', 'BMI']],
palette="coolwarm")
plt.title("□ Distribution of Key Features in Endometriosis Dataset",
fontsize=14)
plt.ylabel("Value")
plt.xticks(rotation=45)
plt.show()
```

```
D:\anaconda\Lib\site-packages\IPython\core\pylabtools.py:170:
UserWarning: Glyph 128202 (\N{BAR CHART}) missing from current font.
  fig.canvas.print_figure(bytes_io, **kw)
```




```
# Checking outliers in BMI using the IQR method
Q1 = df['BMI'].quantile(0.25)
Q2 = df['BMI'].quantile(0.5)
Q3 = df['BMI'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

lower_bound
```

```
12.253932334327025
```

```
upper_bound
```

```
33.78831759692971
```

Age and Chronic Pain level Distributions are quite normal with no major outliers.

BMI has multiple high outliers suggesting few indivuals have High BMI, but they appear to be valid medical cases rather than data errors, so they were retained for analysis.

```
#to check correlation matrix among the numerical values
df.corr()
```

```
                               Age  Menstrual_Irregularity  \
Age                       1.000000               -0.007918
```

```
Menstrual_Irregularity    -0.007918                1.000000
Chronic_Pain_Level        -0.009348                0.000103
Hormone_Level_Abnormality -0.004070                0.014647
Infertility                0.014078               -0.012202
BMI                       -0.011878                0.004868
Diagnosis                 -0.011559                0.095197

                           Chronic_Pain_Level
Hormone_Level_Abnormality  \
Age                                -0.009348                    -
0.004070
Menstrual_Irregularity              0.000103
0.014647
Chronic_Pain_Level                  1.000000
0.002467
Hormone_Level_Abnormality           0.002467
1.000000
Infertility                         0.009299
0.003890
BMI                                 0.005755                    -
0.015499
Diagnosis                           0.116996
0.187039

                          Infertility       BMI  Diagnosis
Age                          0.014078 -0.011878  -0.011559
Menstrual_Irregularity      -0.012202  0.004868   0.095197
Chronic_Pain_Level           0.009299  0.005755   0.116996
Hormone_Level_Abnormality    0.003890 -0.015499   0.187039
Infertility                  1.000000  0.011529   0.096172
BMI                          0.011529  1.000000   0.080310
Diagnosis                    0.096172  0.080310   1.000000
```

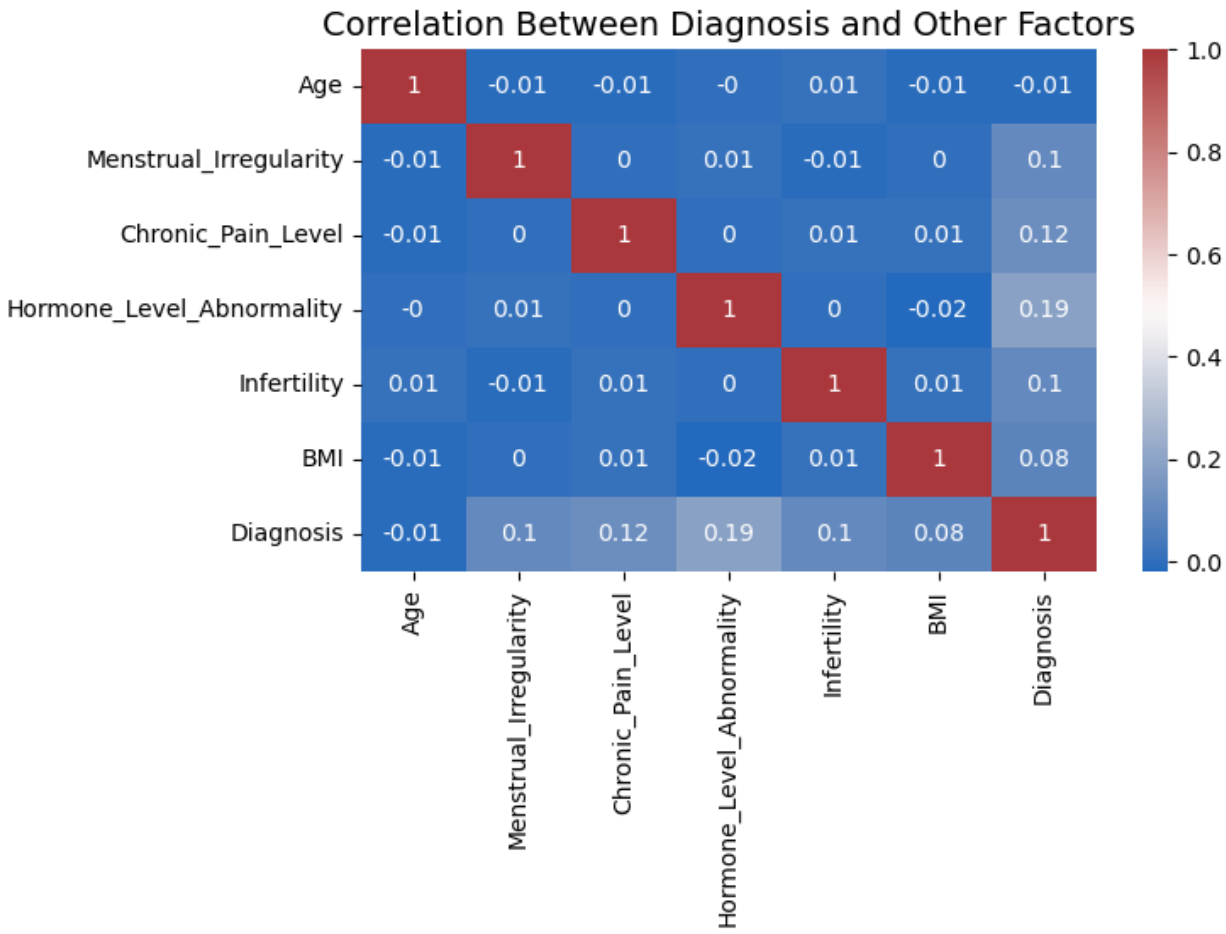
```
#rounding it to 2 decimal place
df.corr().round(2)
```

```
                           Age  Menstrual_Irregularity
Chronic_Pain_Level  \
Age                       1.00                   -0.01
-0.01
Menstrual_Irregularity   -0.01                    1.00
0.00
Chronic_Pain_Level       -0.01                    0.00
1.00
Hormone_Level_Abnormality -0.00                   0.01
0.00
Infertility               0.01                   -0.01
0.01
BMI                      -0.01                    0.00
0.01
```

```
Diagnosis                    -0.01                      0.10
0.12

                             Hormone_Level_Abnormality  Infertility
BMI  \
Age                                              -0.00         0.01 -
0.01
Menstrual_Irregularity                            0.01        -0.01
0.00
Chronic_Pain_Level                                0.00         0.01
0.01
Hormone_Level_Abnormality                         1.00         0.00 -
0.02
Infertility                                       0.00         1.00
0.01
BMI                                              -0.02         0.01
1.00
Diagnosis                                         0.19         0.10
0.08

                             Diagnosis
Age                              -0.01
Menstrual_Irregularity            0.10
Chronic_Pain_Level                0.12
Hormone_Level_Abnormality         0.19
Infertility                       0.10
BMI                               0.08
Diagnosis                         1.00
```

```
#Correlation Between Diagnosis & Other Factors
plt.figure(figsize=(7, 4))
sns.heatmap(df.corr().round(2),annot = True, cmap = 'vlag')
plt.title("Correlation Between Diagnosis and Other Factors", fontsize
= 14)
```

```
Text(0.5, 1.0, 'Correlation Between Diagnosis and Other Factors')
```

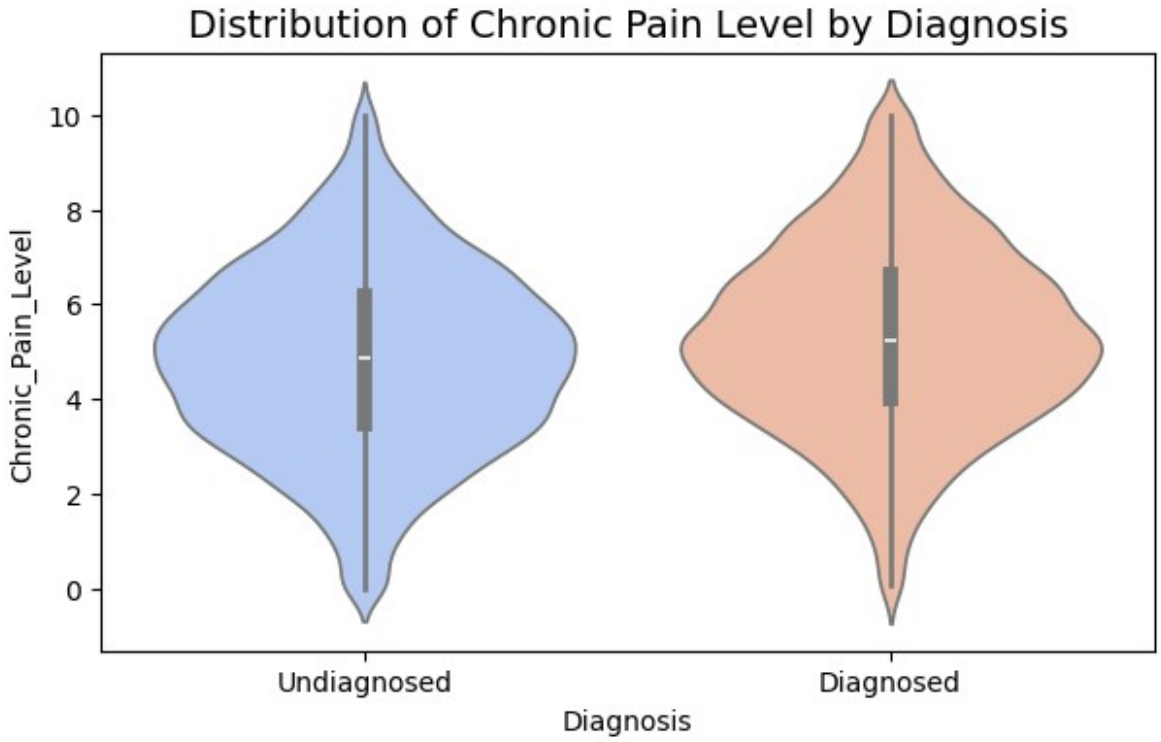


The heatmap shows that Hormone Level Abnormality and Chronic Pain Level have the strongest correlation with endometriosis diagnosis, while BMI and Age have weaker relationships.

```
#Distribution of Chronic Pain level Across Diagnosed and Undiagnosed
Patients
plt.figure(figsize=(7, 4))
sns.violinplot(x="Diagnosis", y="Chronic_Pain_Level", data=df,
palette="coolwarm")
plt.xticks([0, 1], ["Undiagnosed", "Diagnosed"])
plt.title("Distribution of Chronic Pain Level by Diagnosis",
fontsize=14)
plt.show()
```

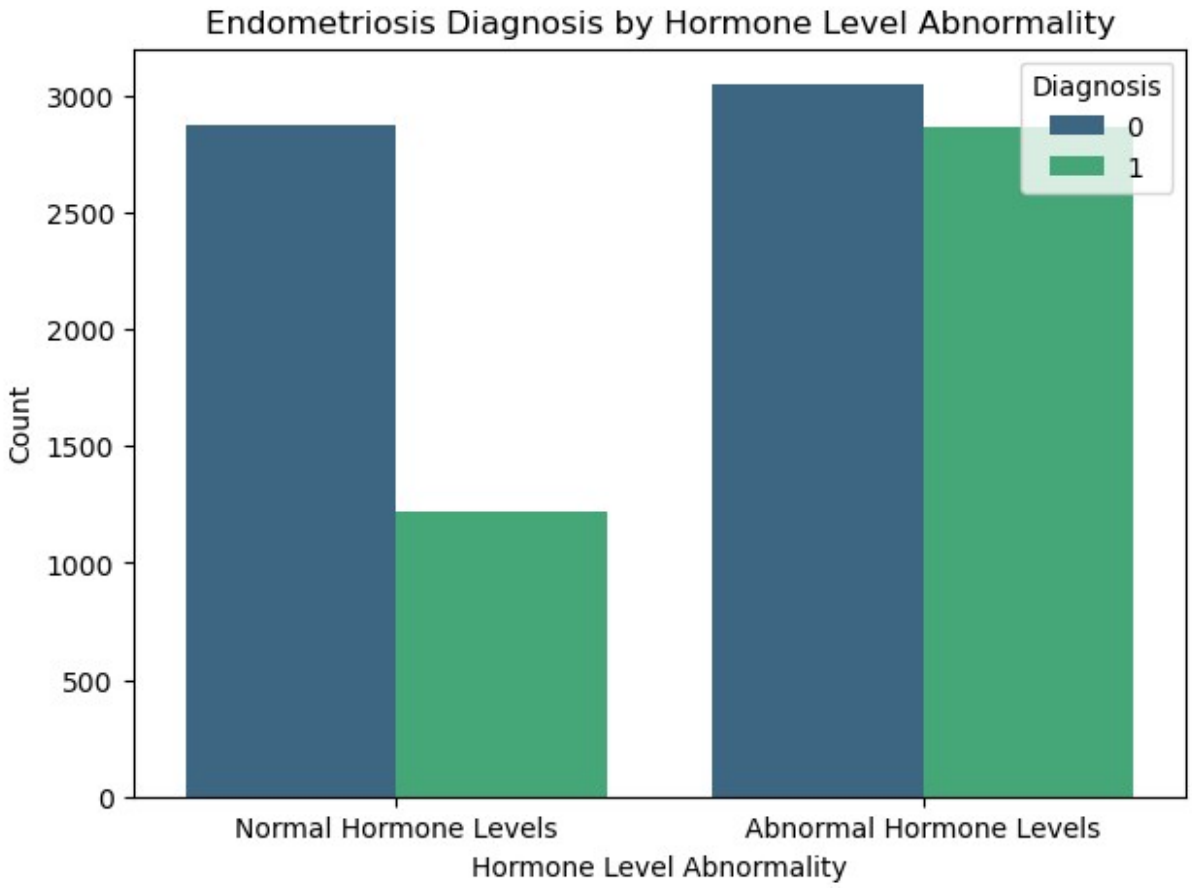
```
C:\Users\KIIT\AppData\Local\Temp\ipykernel_1316\1778860929.py:3:
FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `x` variable to `hue` and set
`legend=False` for the same effect.
```

```
  sns.violinplot(x="Diagnosis", y="Chronic_Pain_Level", data=df,
palette="coolwarm")
```



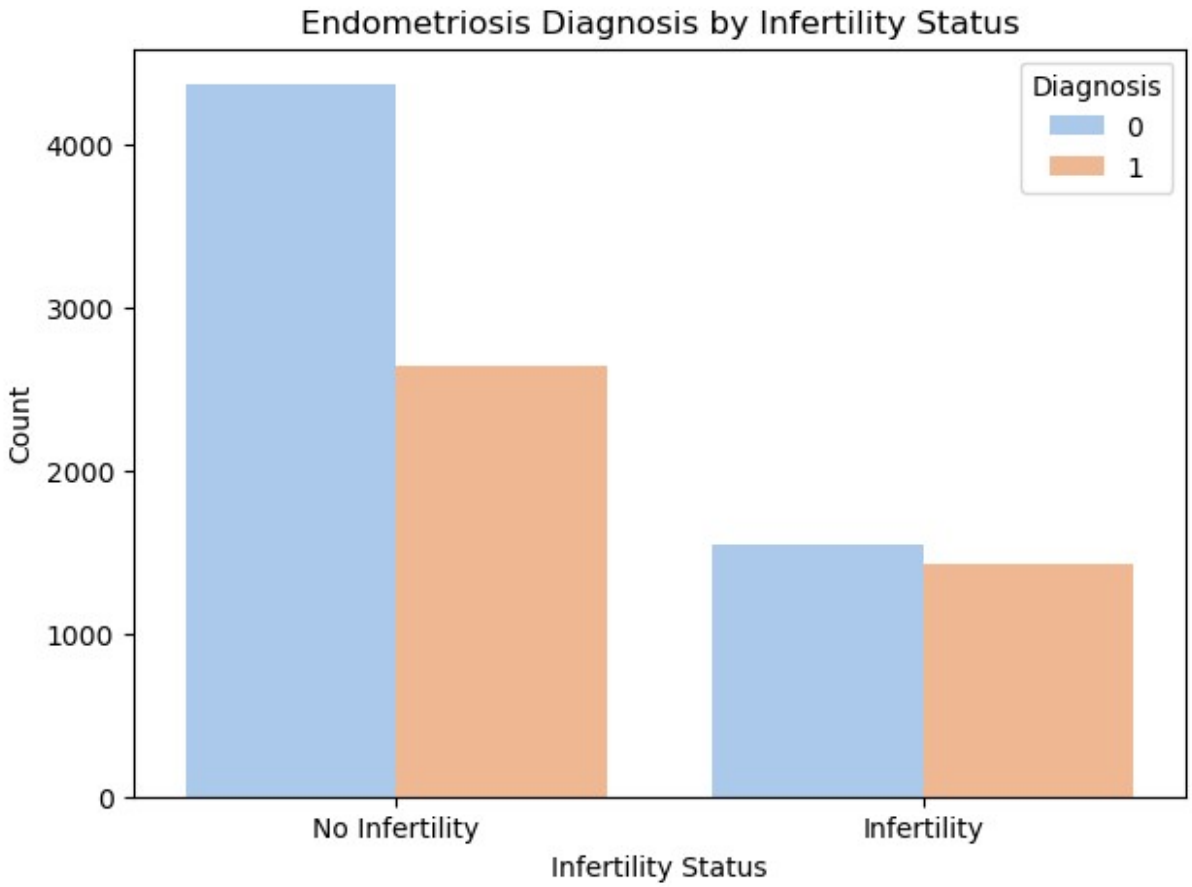
The violin plot shows that patients diagnosed with endometriosis tend to have higher chronic pain levels compared to undiagnosed patients. The wider sections indicate a higher density of cases in that pain range, reinforcing its importance as a key predictor.

```
#Diagnosis by Hormone Level Abnormality
plt.figure(figsize=(7,5))
sns.countplot(x="Hormone_Level_Abnormality", hue="Diagnosis", data=df,
palette="viridis")
plt.xticks([0, 1], ["Normal Hormone Levels", "Abnormal Hormone
Levels"])
plt.xlabel("Hormone Level Abnormality")
plt.ylabel("Count")
plt.title("Endometriosis Diagnosis by Hormone Level Abnormality")
plt.show()
```



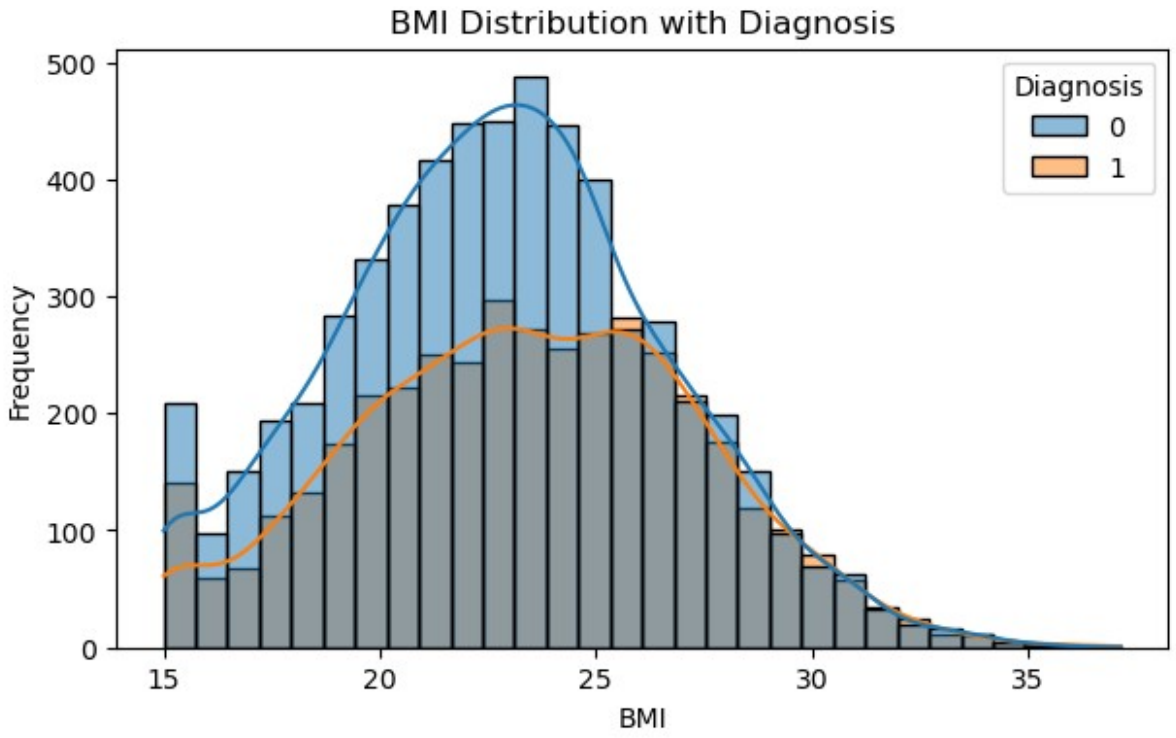
Patients with abnormal hormone levels (1) are more likely to be diagnosed. This confirms hormonal imbalance as a key risk factor for endometriosis.

```
#Diagnosis by Infertility Status
plt.figure(figsize=(7,5))
sns.countplot(x="Infertility", hue="Diagnosis", data=df,
palette="pastel")

# Custom x-axis labels
plt.xticks([0, 1], ["No Infertility", "Infertility"])
plt.xlabel("Infertility Status")
plt.ylabel("Count")
plt.title("Endometriosis Diagnosis by Infertility Status")
plt.show()
```



he bar chart visualizes the relationship between endometriosis diagnosis (0 = No, 1 = Yes) and infertility status. It shows that more individuals without infertility have been diagnosed with endometriosis compared to those with infertility.

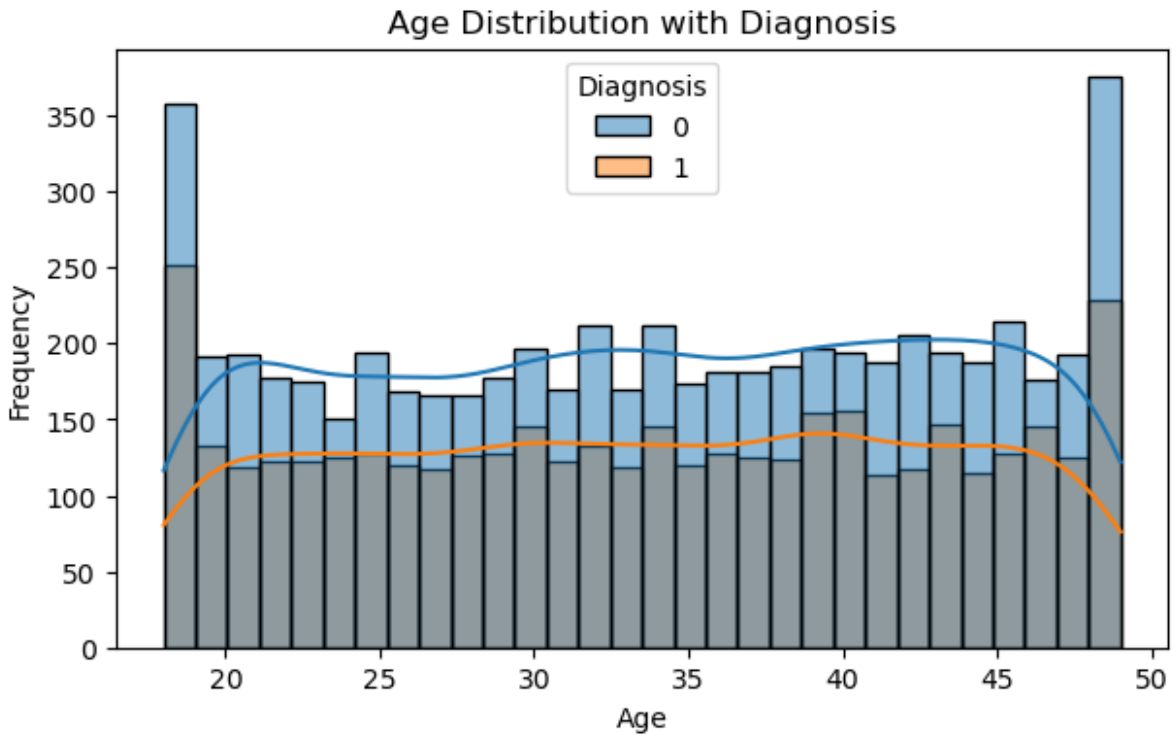
```
#KDE plot for BMI Distribution with Diagnosis
plt.figure(figsize=(7,4))
sns.histplot(df,x="BMI",hue="Diagnosis",kde = True,bins = 30,
color="orange")
plt.title("BMI Distribution with Diagnosis")
plt.ylabel("Frequency")
plt.show()
```



Individuals without the diagnosis (blue) have a higher density in the normal BMI range (18-25), while those with the diagnosis (orange) are more evenly distributed across BMI values.

The density of diagnosed individuals is relatively higher in the overweight and obese BMI range (25+), indicating a possible correlation between higher BMI and the diagnosis.

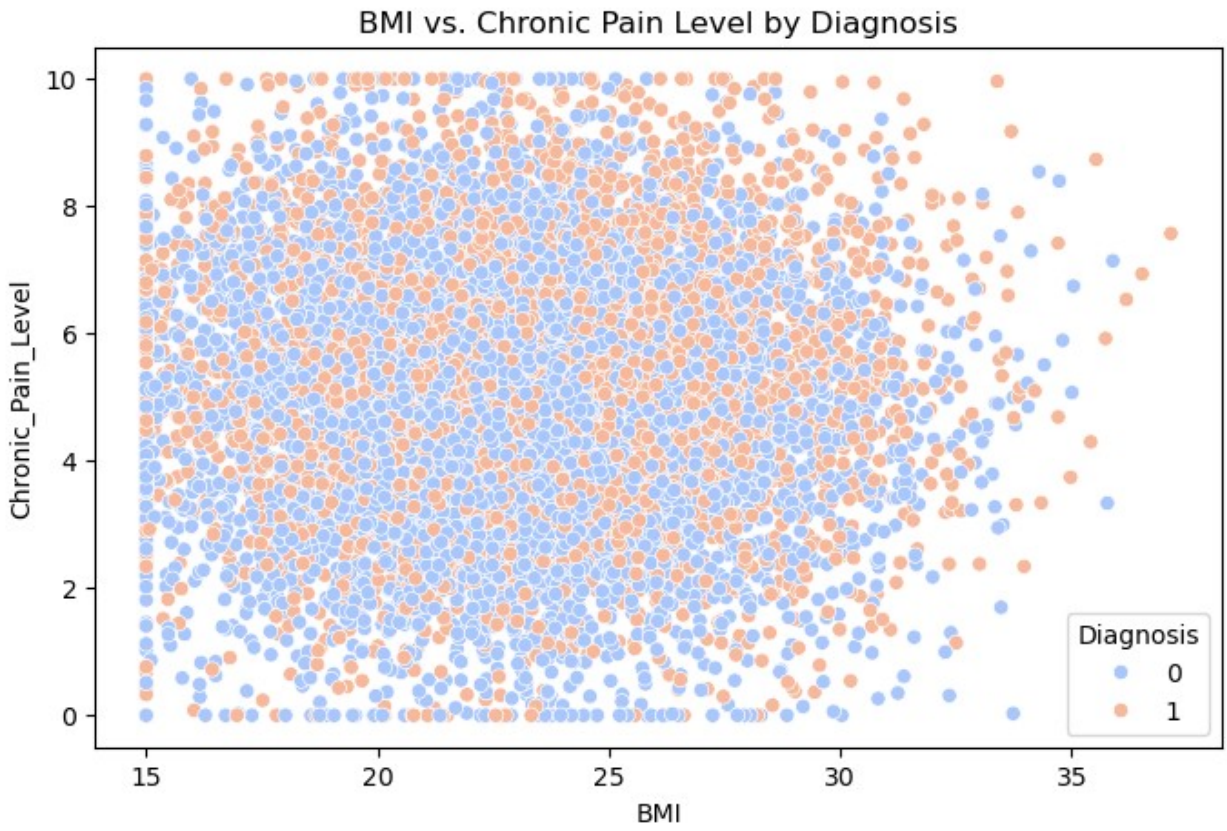
```
#KDE Plot of Age Distribution with Diagnosis
plt.figure(figsize=(7,4))
sns.histplot(df,x="Age", hue="Diagnosis",kde = True,bins = 30,
color="Green")
plt.title("Age Distribution with Diagnosis")
plt.ylabel("Frequency")
plt.show()
```



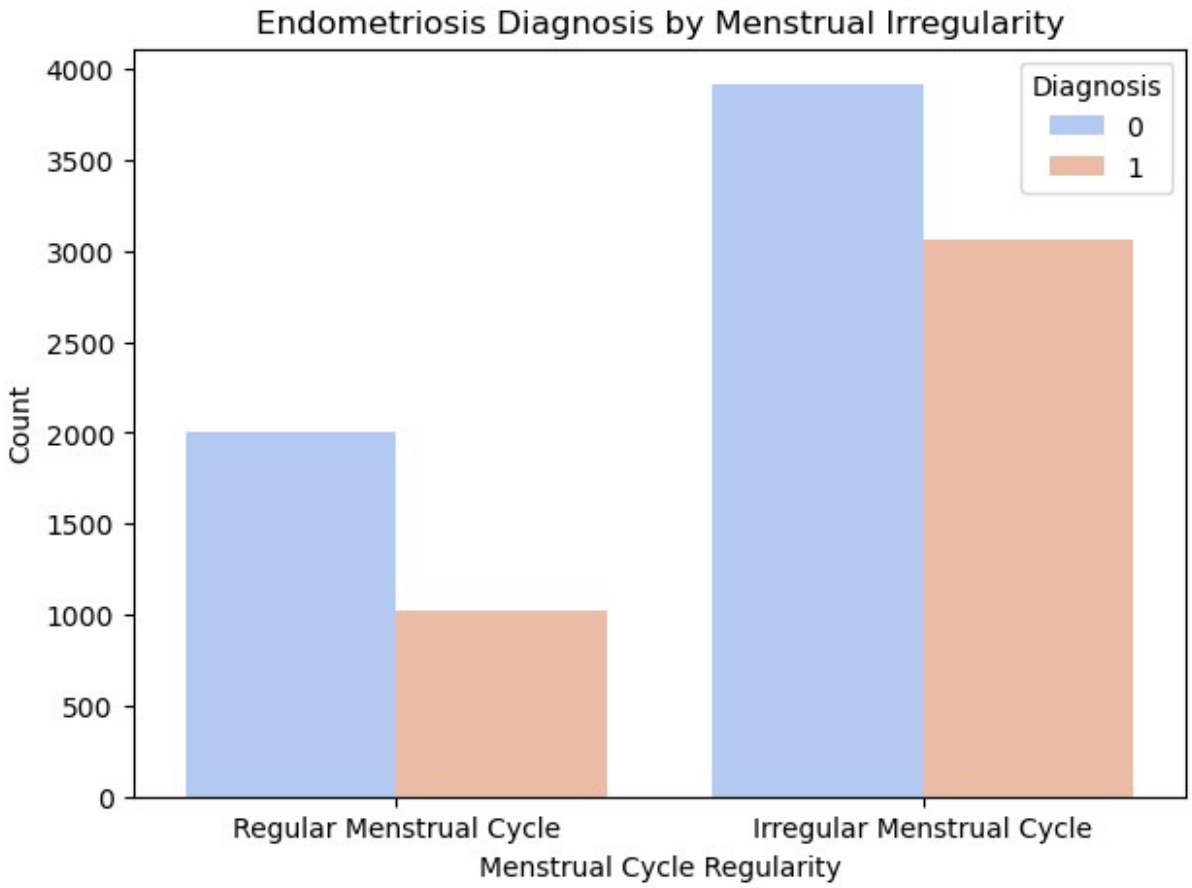
The age distribution of individuals without the diagnosis (blue) is relatively uniform across different age groups, with peaks at the younger (around 20) and older (around 50) age ranges.

Individuals with the diagnosis (orange) show a lower density across all ages, with a slight decline after 45, suggesting that the diagnosis may be less common in older age groups.

```
#To Check if BMI and Chronic Pain Level are linked to Diagnosis
plt.figure(figsize=(8, 5))
sns.scatterplot(x=df["BMI"], y=df["Chronic_Pain_Level"],
hue=df["Diagnosis"], palette="coolwarm")
plt.title("BMI vs. Chronic Pain Level by Diagnosis")
plt.show()
```

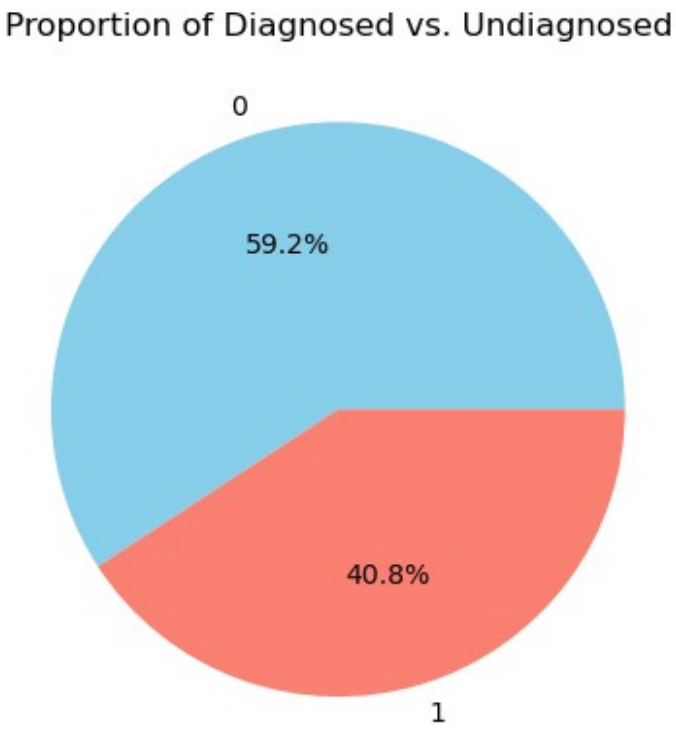



```
#Comparing Diagnosis Frequency for Patients with and without Menstrual Irregularity
plt.figure(figsize=(7,5))
sns.countplot(x="Menstrual_Irregularity", hue="Diagnosis", data=df, palette="coolwarm")
plt.xticks([0, 1], ["Regular Menstrual Cycle", "Irregular Menstrual Cycle"])
plt.xlabel("Menstrual Cycle Regularity")
plt.ylabel("Count")
plt.title("Endometriosis Diagnosis by Menstrual Irregularity")
plt.show()
```



The scatter plot visualizes the relationship between BMI and chronic pain levels, color-coded by endometriosis diagnosis (0 and 1). The data points are densely distributed across the chart, indicating no immediate visible trend between BMI and chronic pain level.

```
#Proportion of Diagnosed vs Undiagnosed
df["Diagnosis"].value_counts().plot.pie(autopct="%1.1f%%",
colors=["skyblue", "salmon"])
plt.title("Proportion of Diagnosed vs. Undiagnosed")
plt.ylabel("")
plt.show()
```



A larger undiagnosed slice suggests that a majority of patients have not been diagnosed.

Key Findings

Hormone Level Abnormality and Chronic Pain Level are the most influential factors in endometriosis diagnosis.

BMI and Menstrual Irregularity have weak correlations with diagnosis.

The dataset contains some outliers, but no extreme multicollinearity issues.

Patients within a certain age range showed more cases, but age itself was not a strong predictor.