BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC KINH TẾ HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN KINH DOANH

ĐỒ ÁN MÔN HỌC

LẬP TRÌNH PHÂN TÍCH DỮ LIỆU

Đề tài:

Phân tích sự ảnh hưởng của các yếu tố đời sống và sức khỏe tới giấc ngủ

Thành viên: Nguyễn Đình Toàn
Nguyễn Thị Thu Trang
Phạm Quốc Thuận
Lưu Trọng Tốt

Giảng viên: TS. Nguyễn An Tế

Thành phố Hồ Chí Minh , ngày 03 tháng 12 năm 2023

LỜI NÓI ĐẦU

Lời đầu tiên, tác giả xin gửi lời cảm ơn đến trường Đại Học UEH đã đưa bộ môn Lập trình phân tích dữ liệu vào trong chương trình giảng dạy. Đặc biệt, tác giả xin trình bày tỏ lòng biết ơn sâu sắc đến thầy Nguyễn An Tế – Giảng viên trường Đại Học UEH, người đã giảng dạy môn Lập trình phân tích dữ liệu cho lớp DS001 một cách tận tình, nhiệt huyết và truyền đạt cho lớp những kiến thức quý báu trong suốt thời gian vừa qua.

Thời gian tham dự lớp học của thầy đã giúp tác giả bổ sung nhiều kiến thức bổ ích, điều đó đã góp phần không nhỏ vào sự thành công của bài tiểu luận cuối kỳ này.

Bộ môn Lập trình phân tích dữ liệu là một môn học thú vị, vô cùng bổ ích đối với mỗi sinh viên ngành Khoa học dữ liệu. Tuy nhiên lượng kiến thức và thời gian còn nhiều hạn chế nên trong quá trình làm bài khó tránh được mắc phải nhiều sai sót, kính mong thầy xem xét và góp ý để giúp bài tiểu luận của tác giả được hoàn thiện hơn.

Xin chân thành cảm ơn!

HCM, ngày 03 tháng 12 năm 2023

Bảng phân công

Thành viên	Phân công	Đánh giá
Phạm Quốc Thuận	Tổng quan đề tài, Phân tích đơn biến, Tinh chỉnh docs, PPT	90%
Nguyễn Đình Toàn	Tổng quan dữ liệu đầu vào, Tiền xử lý, Phân tích đa biến, Kiểm định, Xây dựng mô hình, tinh chỉnh docs	100%
Lưu Trọng Tốt	Tổng quan đề tài, Phân tích đa biến, PPT	60%
Nguyễn Thị Thu Trang	Phân tích đa biến, Kiểm định, nhận xét, tinh chỉnh docs, PPT	100%

CHƯƠNG I. TỔNG QUAN ĐỀ TÀI

1.1 Xác định đề tài

Giấc ngủ đóng một vai trò quan trọng trong việc duy trì sức khỏe tổng thể và tinh thần. Thực tế, giấc ngủ đóng vai trò quan trọng trong việc phục hồi cơ thể, tăng cường hệ miễn dịch và duy trì một tinh thần tốt.

Tuy nhiên, cuộc sống ngày càng bận rộn và áp lực công việc ngày càng gia tăng, khiến cho việc duy trì một chế độ giấc ngủ lành mạnh trở nên khó khăn hơn, dẫn tới việc các bệnh về giấc ngủ ngày càng trở nên phổ biến. Có nhiều yếu tố có thể ảnh hưởng đến giấc ngủ của mọi người, bao gồm căng thẳng, lo lắng, công việc quá tải và thói quen không tốt. Điều này có thể dẫn đến việc mất ngủ, giấc ngủ không đủ hoặc không sâu, và gây ra cảm giác mệt mỏi và khó tập trung trong ngày.

Hiểu được mối quan hệ giữa lối sống và sức khỏe giấc ngủ là điều cần thiết đối với những cá nhân đang tìm cách cải thiện giấc ngủ của mình. Vậy đây là một đề tài thiết thực và thú vị với việc phân tích các yếu tố trên có thể cung cấp những hiểu biết có giá trị về nguyên nhân và tác động của rối loạn giấc ngủ, giúp các cá nhân đưa ra quyết định sáng suốt để tối ưu hóa sức khỏe giấc ngủ của họ.

1.2 Mục đích nghiên cứu

Đề tài tập trung nghiên cứu ảnh hưởng của các yếu tố đến chất lượng giấc ngủ và đưa ra phương hướng cải thiện:

- Tìm hiểu về tầm quan trọng của giấc ngủ đối với sức khỏe và sự phát triển của con người.
- Khám phá các yếu tố ảnh hưởng đến giấc ngủ, bao gồm căng thẳng, lo lắng và thói quen không tốt.
- Đánh giá tác động của việc quản lý căng thẳng và lo lắng trong việc cải thiện giấc ngủ.
- Nghiên cứu về các biện pháp cải thiện giấc ngủ, như tạo môi trường ngủ thoải mái, thiết lập thời gian đi ngủ và thức dậy cố định, và hạn chế sử dụng thiết bị điện tử trước khi đi ngủ.

1.3 Giới hạn phạm vi đề tài

Dự án nghiên cứu tập trung chủ yếu vào ảnh hưởng của các thói quen đối với chất lượng giấc ngủ và đề xuất phương hướng cải thiện. Tuy nhiên, có một số giới hạn phạm vi cần được xác định:

- **Phạm vi Nhóm Đối Tượng:** Nghiên cứu sẽ tập trung chủ yếu vào người trưởng thành và có thể có giới hạn trong một đối tượng cụ thể như người làm việc văn phòng, không bao quát mọi đối tượng như trẻ em hoặc người già.
- **Thời Gian Nghiên Cứu:** Phạm vi thời gian nghiên cứu có thể bị hạn chế, và sẽ tập trung vào một khoảng thời gian cụ thể để thu thập dữ liệu và phân tích ảnh hưởng của thói quen đối với giấc ngủ.
- **Yếu Tố Ngoại Vi:** Mặc dù sẽ xem xét một số yếu tố như căng thẳng và lo lắng, nhưng không thể bao quát toàn bộ các yếu tố có thể ảnh hưởng đến giấc ngủ, và một số yếu tố ngoại vi có thể không được đề cập đến.

- **Giải Pháp Cải Thiện:** Phương hướng cải thiện giấc ngủ sẽ được đề xuất dựa trên những yếu tố được nghiên cứu, nhưng không đi sâu vào các liệu pháp y tế cụ thể hoặc tình huống y tế phức tạp.

1.4 Phương pháp nghiên cứu

- EDA: Sử dụng các biểu đồ thể hiện sự tương quan cũng như làm rõ mục đích nghiên cứu đề tài, sự liên kết với nhau giữa các biến.
- Trực quan hóa dữ liệu: Sử dụng các loại biểu đồ chuyên dụng và phù hợp với mục đích trực quan hoá các dữ liệu, giúp người đọc báo cáo dễ dàng quan sát và đánh giá kết quả phân tích.
- Các loại biểu đồ: bar chart, pie chart,...
- Kiểm định Chi Square: Kiểm định tính độc lập giữa 2 biến phân loại, xác định xem liệu có mối liên hệ giữa 2 biến phân loại hay không.
- Kiểm định Anova: Kiểm định giá trị trung bình của một biến liên tục giữa ba hoặc nhiều nhóm độc lập.
- Kiểm định T-test: Kiểm định giá trị trung bình của một biến liên tục giữa hai nhóm độc lập.
- Mô hình dự báo: KNN, Random Forest, Linear,...

1.5 Ngôn ngữ sử dụng

Ngôn ngữ lập trình: Python.

CHƯƠNG II. TỔNG QUAN BỘ DỮ LIỆU NGHIÊN CỨU

2.1 Tổng quan bộ dữ liệu thu thập

2.1.1 Giới thiệu bộ dữ liệu

Bộ dữ liệu gốc Sleep Health and Lifestyle chứa các thông tin liên quan tới giấc ngủ và thói quen hàng ngày. Bộ dữ liệu có tổng cộng 13 thuộc tính và 374 bản ghi được ghi nhận.

2.1.2 Các thuộc tính của bộ dữ liệu

STT	Tên thuộc tính	Giải thích	Giá trị	Ghi chú
1	Person ID	Mã định danh	374 giá trị	Mỗi mã định danh chỉ thuộc về một người.
2	Age	Độ tuổi	27 - 59	
3	Gender	Giới tính	- Female *(Nữ)* - Male *(Nam)*	
4	Occupation	Nghề nghiệp	11 giá trị	
5	Sleep duration	Số giờ ngủ trong một ngày (Đơn vị: giờ)	5.8 - 8.5	
6	Quality of Sleep	Chất lượng giấc ngủ	4 - 9	Cấp độ từ 4 - 9 thể hiện mức độ chất lượng giấc ngủ từ thấp tới cao.
6	Physical Activity Level	Thời gian hoạt động thể chất trong ngày (Đơn vị: phút/ giờ)	30 - 90	
7	Stress Level	Mức độ căng thẳng	1 - 10	Các cấp độ từ 1 - 10 thể hiện mức độ căng thẳng từ ít tới nhiều.
8	BMI category	Chỉ số BMI	- Normal *(Bình thường)* - Overweight *(Thừa cân)* - Obese *(Béo phì)*	

9	Blood Pressure	Chỉ số huyết áp (Huyết áp tâm thu/ Huyết áp tâm trương)		
10	Heart Rate	Nhịp tim (Đơn vị: bpm)	65 - 86	
11	Daily Steps	Số bước chân mỗi ngày	20 giá trị	
12	Sleep Disorder	Các bệnh về giấc ngủ	- Insomnia *(Chứng mất ngủ)* - Sleep Apnea *(Chứng ngưng thở)* - None *(Không có)*	

2.2 Điều chỉnh bộ dữ liệu

- Tổng quan bộ dữ liệu:

```
df.info()
Output:
RangeIndex: 374 entries, 0 to 373
Data columns (total 13 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Person ID                374 non-null    int64
 1   Gender                   374 non-null    object
 2   Age                      374 non-null    int64
 3   Occupation               374 non-null    object
 4   Sleep Duration           374 non-null    float64
 5   Quality of Sleep         374 non-null    int64
 6   Physical Activity Level  374 non-null    int64
 7   Stress Level             374 non-null    int64
 8   BMI Category             374 non-null    object
 9   Blood Pressure           374 non-null    object
 10  Heart Rate               374 non-null    int64
 11  Daily Steps              374 non-null    int64
 12  Sleep Disorder           374 non-null    object
dtypes: float64(1), int64(7), object(5)
memory usage: 38.1+ KB
```

- Thêm cột Awakening: Mục tiêu của bộ dữ liệu này là xem xét các thông số của người bị bệnh về giấc ngủ "Sleep Disorder", với 3 trường hợp: những người bình thường 'None', những người bị bệnh mất ngủ 'Insomnia' và những người bị bệnh ngưng thở lúc ngủ 'Sleep Apnea'. Vì vậy, nhóm quyết định tạo thêm một cột liên quan tới các bệnh lý này, đó là số lần thức giấc trong giấc ngủ "Awakening'

```
# Tạo điều kiện cho np.select
conditions = [
    (df['Sleep Disorder'] == 'Sleep Apnea'),
    (df['Sleep Disorder'] == 'Insomnia'),
    (df['Stress Level'] > 6)
]
# Tạo giá trị tương ứng cho từng điều kiện
values = [
    np.floor(np.random.uniform(3, 6, len(df))),
    np.floor(np.random.uniform(1, 3, len(df))),
    np.floor(np.random.uniform(0, 2, len(df)))
]
# Mặc định cho các trường hợp khác
default_value = np.floor(np.random.uniform(0, 1, len(df)))
# Thêm cột 'Awakening' dựa trên điều kiện
df['Awakening'] = np.select(conditions, values, default_value)
```

  - Với những người bị ngưng thở lúc ngủ 'Sleep Apnea', dù không có kết quả nghiên cứu thực tế về số lần thức giấc của bệnh này, nhóm sẽ cho cao hơn 2 trường hợp còn lại. Do đó, nhóm cho ngẫu nhiên số lần thức giấc từ 3 tới 6.
  - Với những người bị mất ngủ 'Insomnia' thường bị thức giấc vì tiếng ồn, vì ánh sáng,... thì nhóm sẽ cho ngẫu nhiên số lần thức giấc từ 1 tới 3 lần.
  - Với những người bình thường 'None', nếu mức độ stress cao trên 6 thì nhóm sẽ cho ngẫu nhiên từ 0 tới 2 lần, còn lại sẽ từ 0 tới 1 lần.
    → Hãy xem cột mới này có đóng góp gì cho bộ dữ liệu hay không.
- Thêm ngẫu nhiên dữ liệu bị thiếu: Ở đây bộ dữ liệu xem xét thoáng qua là một bộ dữ liệu rất đầy đủ. Vì vậy để có thêm công việc xử lý, nhóm quyết định tạo hàm để tạo missing value một cách ngẫu nhiên từ 1% tới 10% cho ngẫu nhiên từ 1 đến 3 cột.

```
num_columns_to_make_missing = int(np.round(np.random.uniform(1, 3)))
columns_to_make_missing = np.random.choice(df.columns,
num_columns_to_make_missing, replace=False)

# Tạo giá trị thiếu trong mỗi cột, trừ 'Sleep disorder'
for column in columns_to_make_missing:
    if column != 'Sleep Disorder':
```

```
        column_size = len(df[column])
        missing_rate = np.random.uniform(0.01, 0.10)
        num_missing_values = int(column_size * missing_rate)
        missing_indices = np.random.choice(df.index,
num_missing_values, replace=False)
        df.loc[missing_indices, column] = np.nan

print(df)
```

Các thay đổi trên được thực hiện để bổ sung thông tin vào bộ dữ liệu gốc, mang lại cơ hội khám phá những hướng tiếp cận mới trong quá trình xử lý dữ liệu; đồng thời làm cho việc biểu diễn dữ liệu trở nên thuận lợi hơn cho quá trình mô hình hóa, phân tích tương quan và trực quan hóa dữ liệu.

Việc thêm dữ liệu mới không chỉ làm giàu nguồn thông tin mà còn mở rộng cơ hội để thực hiện những phân tích chi tiết hơn. Bên cạnh đó, việc tạo ra giá trị thiếu ngẫu nhiên cung cấp một góc nhìn động lực cho quá trình nghiên cứu, khiến cho nhóm có thể đối mặt với những thách thức và tìm kiếm các phương hướng xử lý sáng tạo. Điều này không chỉ tạo ra sự đa dạng trong quy trình nghiên cứu mà còn khuyến khích sự sáng tạo và sự linh hoạt trong việc đưa ra các giải pháp.

CHƯƠNG III. TIỀN XỬ LÝ

3.1 Kiểm tra tình trạng bộ dữ liệu gốc:

3.1.1 Tình trạng bộ dữ liệu thu thập:

```
df.head()
```

Output:

	Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder	Awakening
0	1	Male	27	Software Engineer	6.1	6	42	6	Overweight	126/83	77	4200	None	0.0
1	2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None	1.0
2	3	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None	0.0
3	4	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea	4.0
4	5	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea	4.0

```
df.isnull().sum()
missing_data_summary=pd.DataFrame({'Missing Count':
df.isnull().sum(),
                                   'Missing Percentage':
df.isnull().mean()*100})
print(missing_data_summary)
```

Output:

	Missing Count	Missing Percentage
Person ID	0	0.000000
Gender	0	0.000000
Age	0	0.000000
Occupation	0	0.000000
Sleep Duration	0	0.000000
Quality of Sleep	0	0.000000
Physical Activity Level	0	0.000000
Stress Level	0	0.000000
BMI Category	0	0.000000
Blood Pressure	0	0.000000
Heart Rate	**37**	**9.893048**
Daily Steps	0	0.000000
Sleep Disorder	0	0.000000
Awakening	**25**	**6.684492**

Dựa trên tổng quan về bộ dữ liệu, nhóm nhận thấy có hai vấn đề chính ở khâu tiền xử lý là: giá trị thiếu (missing values) và những cột có giá trị hiển thị không phù hợp, do đó cần được xử lý để bộ dữ liệu trở nên linh hoạt và có tính nhất quán hơn.

- Giá trị thiếu (missing values):

Có thể cần thực hiện kiểm tra chi tiết để xác định phạm vi và số lượng giá trị thiếu trong từng cột. Các phương pháp xử lý giá trị thiếu có thể bao gồm: điền giá trị trung bình, trung vị hoặc sử dụng các phương pháp máy học phức tạp hơn để dự đoán giá trị thiếu.

- Các cột với giá trị hiển thị chưa hợp lý hoặc không đóng góp ý nghĩa:
  Một số cột có thể chứa giá trị không phù hợp hoặc không rõ ràng. Việc này có thể bao gồm các giá trị ngoại lệ (outliers) hoặc các giá trị không hợp lý đối với loại dữ liệu được mô tả. Các giá trị này cần được kiểm tra và điều chỉnh để đảm bảo tính chính xác và đồng nhất của bộ dữ liệu.

Quá trình xử lý dữ liệu là một phần quan trọng trong quá trình chuẩn bị dữ liệu cho phân tích và mô hình hóa. Việc giải quyết những vấn đề này sẽ giúp bộ dữ liệu trở nên đáng tin cậy hơn và cung cấp nền tảng tốt cho các phân tích và nghiên cứu sau này.

3.2 Xử lý dữ liệu

3.2.1 Chỉnh dạng dữ liệu

a. Xóa cột:

Đầu tiên có thể thấy được cột "Person ID" có thể xem xét loại bỏ khỏi bộ dữ liệu. Điều này dựa trên nhận định rằng ngoài việc cung cấp thông tin về thứ tự người khảo sát, cột này không mang lại giá trị thống kê đáng kể cho quá trình phân tích và xây dựng mô hình sau này. Và ngoài ra việc loại bỏ cũng giúp giảm thiểu lượng dữ liệu cần xử lý, tránh nhiễu không cần thiết và tối ưu hóa quá trình nghiên cứu. Vì vậy, nhóm sẽ quyết định loại bỏ cột này ra khỏi bộ dữ liệu.

```
# Vì cột này chỉ mang tính chất đánh số thứ tự không mang ý nghĩa
thống kê => Drop
df = df.drop(['Person ID'], axis=1)
# Cập nhật lại biến numerical
numeric_vars = df.select_dtypes(exclude=['object']).columns
print("\nBiến numerical:")
print(numeric_vars)
```

Output:

```
Biến numerical:
Index(['Person ID', 'Age', 'Sleep Duration', 'Quality of Sleep',
       'Physical Activity Level', 'Stress Level', 'Heart Rate',
'Daily Steps',
       'Awakening'],
      dtype='object')
```

b. Chỉnh cột:

Bước tiếp theo là chỉnh những cột có giá trị chưa hợp lý cho việc phân tích và ở đây là biến "Blood Pressure" khi giá trị của cột này ở dạng tỉ số (giữa huyết áp tâm thu và huyết áp tâm trương). Như ví dụ dưới đây:

```
df['Blood Pressure'].describe()
```

Output:

```
count         374
unique         25
top        130/85
freq           99
Name: Blood Pressure, dtype: object
```

Vậy ta sẽ tách biến "Blood Pressure" thành hai cột riêng lẻ, một là huyết áp tâm thu (Blood Pressure 1) và một là huyết áp tâm trương (Blood Pressure 2), điều này sẽ mang lại một số lợi ích:

- Dễ hiểu hơn: Việc tách biến giúp làm cho dữ liệu trở nên dễ hiểu hơn, vì nó phản ánh rõ ràng hai thành phần chính của "Blood Pressure". Các cột riêng lẻ giúp làm cho tập dữ liệu trở nên có ý nghĩa hơn trong quá trình phân tích.
- Thuận tiện cho phân tích: Có thể dễ dàng áp dụng các phương pháp phân tích dựa trên các biến độc lập (univariate analysis) hoặc so sánh giữa các nhóm (group comparison) cho mỗi thành phần riêng biệt mà không cần xử lý giá trị tỉ lệ.
- Thuận tiện cho quá trình xử lý dữ liệu: Việc tách biến giúp làm cho việc xử lý và làm sạch dữ liệu trở nên thuận tiện hơn, đặc biệt là khi cần xử lý giá trị không hợp lý hoặc thiếu sót.

Với việc tách biến này, nhóm có thể thực hiện phân tích chi tiết và sâu sắc hơn về huyết áp trong tập dữ liệu.

```
df[['Blood Pressure 1', 'Blood Pressure 2']] = df['Blood
Pressure'].str.split('/', expand=True).astype(int)
df.drop(columns='Blood Pressure', inplace=True)
df.head()
```

Output:

	Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Heart Rate	Daily Steps	Sleep Disorder	Awakening	Blood Pressure 1	Blood Pressure 2
0	1	Male	27	Software Engineer	6.1	6	42	6	Overweight	77	4200	None	0.0	126	83
1	2	Male	28	Doctor	6.2	6	60	8	Normal	75	10000	None	1.0	125	80
2	3	Male	28	Doctor	6.2	6	60	8	Normal	75	10000	None	0.0	125	80
3	4	Male	28	Sales Representative	5.9	4	30	8	Obese	85	3000	Sleep Apnea	4.0	140	90
4	5	Male	28	Sales Representative	5.9	4	30	8	Obese	85	3000	Sleep Apnea	4.0	140	90

Tương tự với biến BMI Category, giá trị Normal Weight và Normal là gần như không khác biệt về mặt ý nghĩa, nên ta sẽ gộp 2 giá trị này lại thành ‘Normal’ để tiện biểu diễn trực quan và xử lý sau này:

```
df['BMI Category'].unique()
```

Output:

```
array(['Overweight', 'Normal', 'Obese', 'Normal Weight'], dtype=object)
```

```
# Thay thế Normal Weight to Normal
df['BMI Category'] = df['BMI Category'].replace('Normal Weight',
'Normal')
df['BMI Category'].unique()
```

Output:

```
array(['Overweight', 'Normal', 'Obese'], dtype=object)
```

3.2.2 Xử lý missing value

```
df.isnull().sum()
missing_data_summary=pd.DataFrame({'Missing Count':
df.isnull().sum(),'Missing Percentage': df.isnull().mean()*100})
print(missing_data_summary)
```

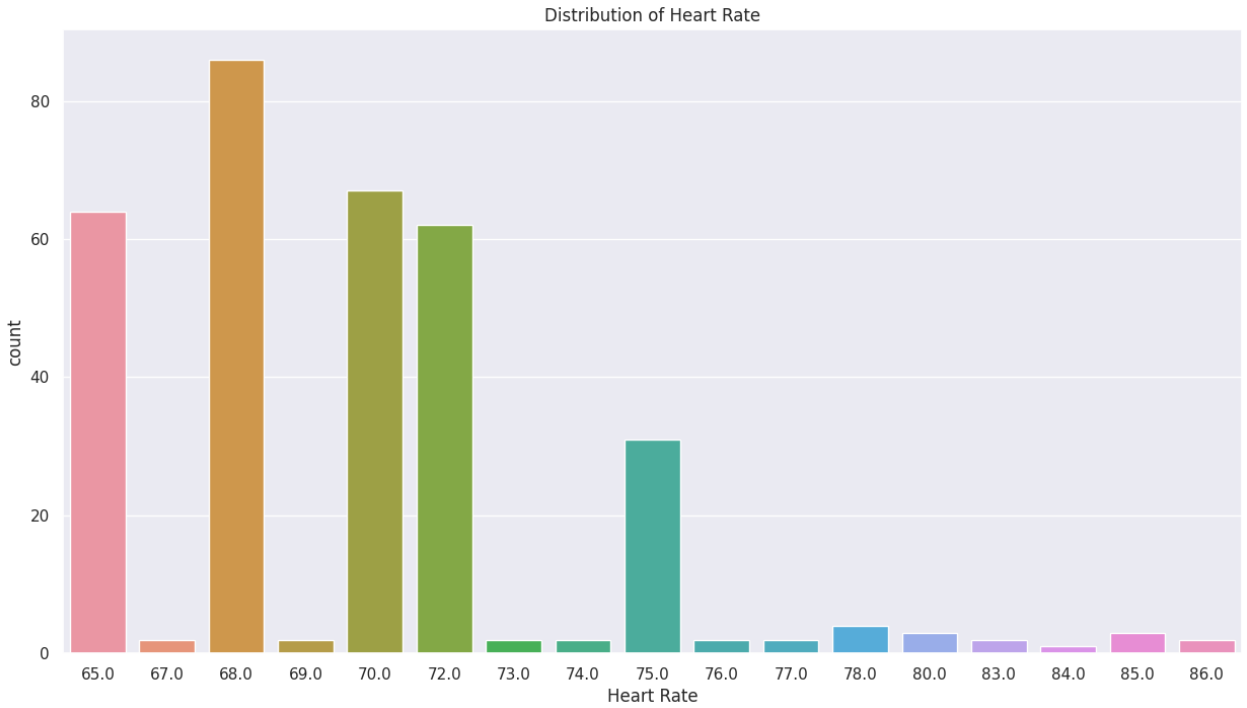
Output:

```
                         Missing Count  Missing Percentage
Person ID                            0            0.000000
Gender                               0            0.000000
Age                                  0            0.000000
Occupation                           0            0.000000
Sleep Duration                       0            0.000000
Quality of Sleep                     0            0.000000
Physical Activity Level              0            0.000000
Stress Level                         0            0.000000
BMI Category                         0            0.000000
Heart Rate                          37            9.893048
Daily Steps                          0            0.000000
Sleep Disorder                       0            0.000000
Awakening                           25            6.684492
Blood Pressure 1                     0            0.000000
Blood Pressure 2                     0            0.000000
```

Ở đây tác giả thấy được rằng có 2 cột chứa giá trị Missing value là “Heart Rate” và “Awakening”.

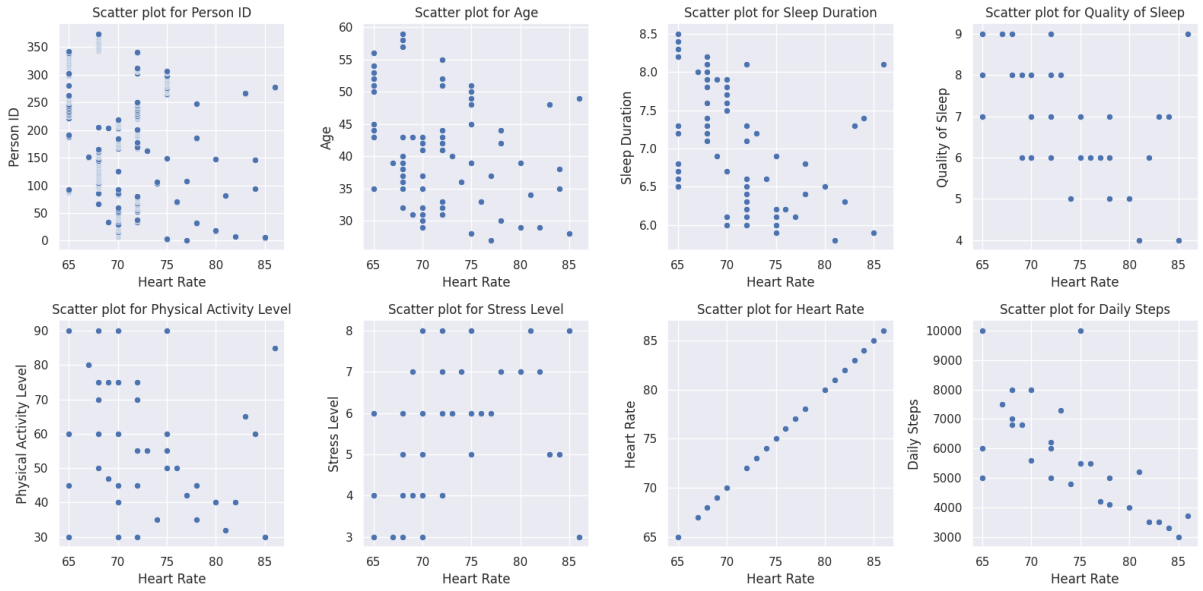
a. Biến “Heart Rate”:

Trước tiên hãy xem qua biểu đồ phân phối của biến này:



Các miền giá trị của biến "Heart Rate" mặc dù có xu hướng lệch trái nhưng tỉ lệ phân bổ không hoàn toàn đồng đều khi phần lớn các giá trị tập trung chủ yếu từ 65 tới 75. Vậy hãy xem có sự tương quan giữa biến này với các biến khác hay không để có thể làm tiền để bổ sung các giá trị bị thiếu:

```
cols_in_row = len(numeric_vars)
fig, axes = plt.subplots(2, cols_in_row // 2, figsize=((2 *
cols_in_row, 8)))
axes = axes.flatten()
for i, numeric_var in enumerate(numeric_vars):
    sns.scatterplot(data=df, x='Heart Rate', y=numeric_var,
ax=axes[i])
    axes[i].set_title(f'Scatter plot for {numeric_var}')
plt.tight_layout()
plt.show()
```

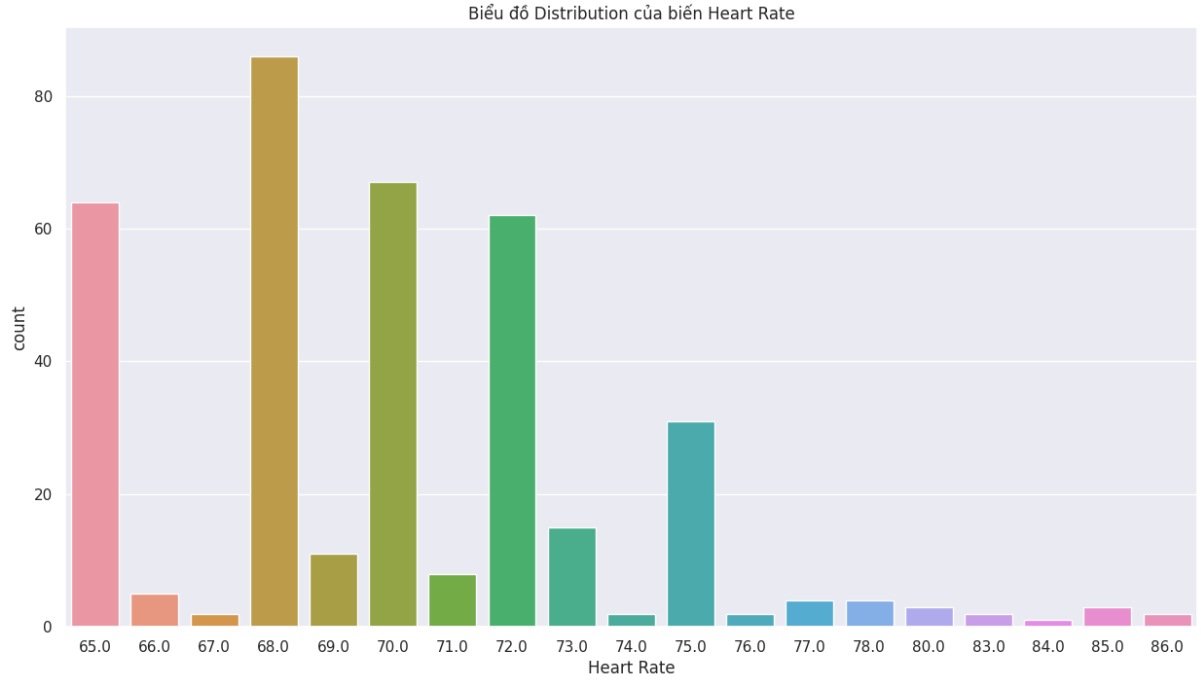


Có thể thấy được có 2 biến có sự tương quan với "Heart Rate" ở đây là "Quality of Sleep" và "Daily Steps". Vậy qua 2 biến này hãy xây dựng mô hình hồi quy dùng Linear Regression để bổ sung các giá trị bị thiếu cho "Heart Rate":

```
from sklearn.linear_model import LinearRegression

# Tách dữ liệu thành hai phần: một phần chứa missing value và phần
còn lại
df_missing = df[df['Heart Rate'].isnull()]
df_not_missing = df.dropna(subset=['Heart Rate', 'Daily Steps',
'Quality of Sleep'])

# Chia dữ liệu có giá trị thành features (X) và target (y)
X_train = df_not_missing[['Daily Steps', 'Quality of Sleep']]
y_train = df_not_missing['Heart Rate']

# Sử dụng mô hình hồi quy tuyến tính để dự đoán giá trị missing
model = LinearRegression()
model.fit(X_train, y_train)
predicted_values = model.predict(df_missing[['Daily Steps', 'Quality
of Sleep']])

# Làm tròn giá trị dự đoán thành số nguyên
predicted_values_rounded = np.round(predicted_values).astype(int)
df.loc[df['Heart Rate'].isnull(), 'Heart Rate'] =
predicted_values_rounded
```

- Giải thích:
    - Tác giả sẽ chia dữ liệu của "Heart Rate" thành 2 phần, một phần không chứa missing value để train và phần còn lại chứa missing value để điền vào sau khi thuật toán đã học được những sự tương quan giữa các biến.
    - Ở đây sẽ lấy phần nguyên của giá trị được dự đoán (round) vì các chỉ số nhịp tim đều là giá trị nguyên.

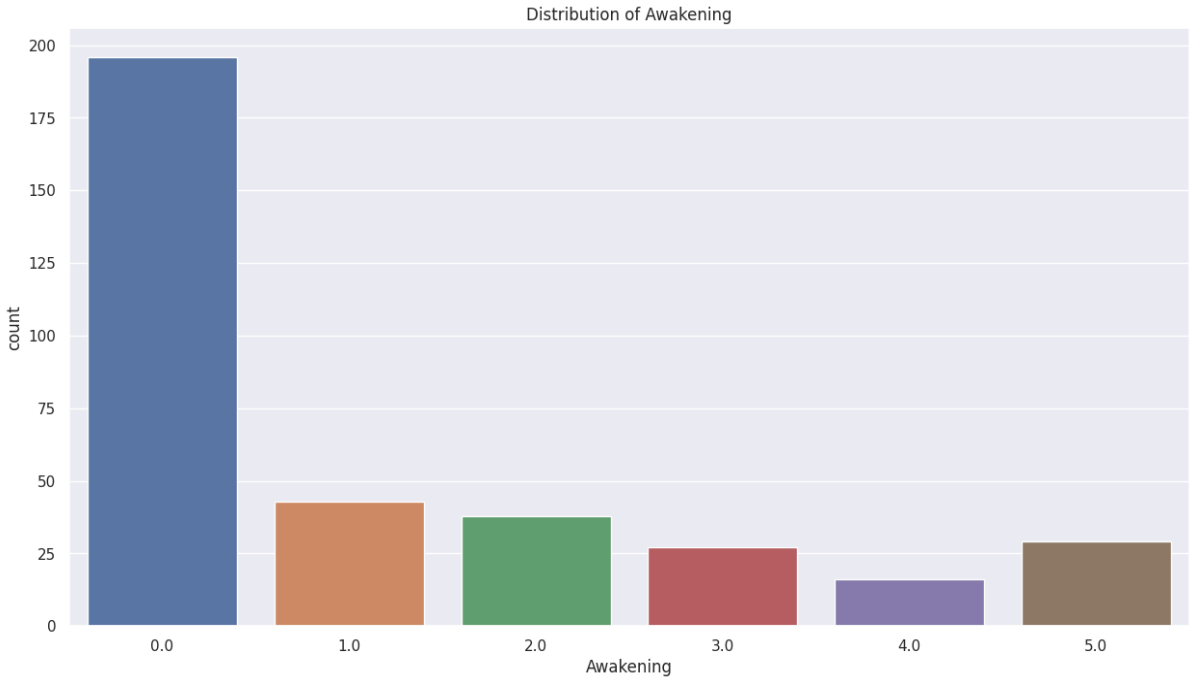
Kết quả sau khi chạy thuật toán:



- Đánh giá: Có thể thấy sau khi bổ sung giá trị thiếu bằng thuật toán hồi quy thì tổng quan phân phối của biến "Heart Rate" cũng không thay đổi nhiều, biểu đồ vẫn lệch trái và các giá trị xuất hiện phổ biến vẫn là 68 hay 70.

b. Biến "Awakening":

Trước tiên hãy cùng xem qua phân phối của biến này:



Có thể thấy được giá trị xuất hiện nhiều nhất là 0, kiểm định đó thấy được những người trong tập khảo sát không có tình trạng thức giấc nhiều lần trong đêm. Và dựa theo biểu đồ này, tác giả sẽ điền giá trị bị thiếu bằng giá trị xuất hiện nhiều nhất là 0.0

```
df['Awakening'].fillna(df['Awakening'].mode()[0], inplace=True)
```

Kiểm tra lại missing value của bộ dữ liệu:

```
df.isnull().sum()
missing_data_summary = pd.DataFrame({'Missing Count':
df.isnull().sum()
                       , 'Missing Percentage': df.isnull().mean() *
100})
print(missing_data_summary)
```

Output:

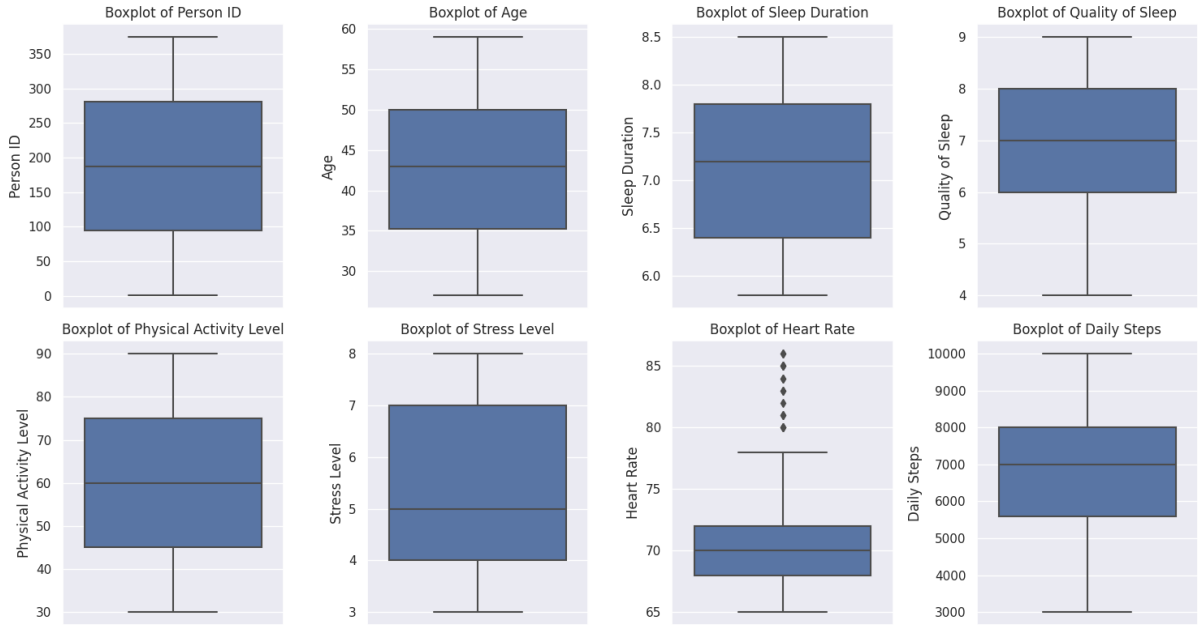
```
                        Missing Count  Missing Percentage
Person ID                           0            0.000000
Gender                              0            0.000000
Age                                 0            0.000000
Occupation                          0            0.000000
Sleep Duration                      0            0.000000
Quality of Sleep                    0            0.000000
Physical Activity Level             0            0.000000
Stress Level                        0            0.000000
BMI Category                        0            0.000000
Heart Rate                          0            0.000000
Daily Steps                         0            0.000000
Sleep Disorder                      0            0.000000
Awakening                           0            0.000000
```

Blood Pressure 1	0	0.000000
Blood Pressure 2	0	0.000000

Vậy là bộ dữ liệu đã không còn chứa Missing Value. Điều này là một tiền đề tốt để đến những phần xử lý tiếp theo cho bộ dữ liệu:

3.2.3 Lọc giá trị nhiễu Outliers Data:

Nhóm dùng boxplot để kiểm tra outliers của bộ dữ liệu:



Trong bộ dữ liệu, có một biến số định lượng duy nhất chứa giá trị ngoại lai là "Heart Rate" khi có các giá trị ngoài 80 như trên boxplot đã thể hiện. Tuy nhiên, nếu quan sát theo ngữ cảnh của biến này là chỉ số nhịp tim thì việc có những giá trị trong khoảng 80-85 không phải là không thể dữ lại vì đó là những chỉ số hoàn toàn có ý nghĩa. Mặc dù vậy để xem xét có nên xử lý những giá trị này hay không thì nhóm sẽ kiểm tra mức độ ảnh hưởng của những giá trị ngoại lai này đến khả năng phát hiện bệnh về giấc ngủ.

- Mục tiêu của kiểm định này là xác định xem nhóm người có nhịp tim vượt quá mức 78 có xu hướng phát hiện bệnh về giấc ngủ nhiều hơn so với nhóm còn lại hay không.
- Nếu kết quả không chứng minh điều này, tác giả có thể xem xét việc xử lý những giá trị ngoại lai này, đặc biệt là khi một số mô hình hồi quy có thể phản ứng nhạy cảm với sự hiện diện của chúng. Ngược lại, nếu có chứng cứ chứng minh giá trị thống kê của nhóm có nhịp tim cao, thì việc giữ lại thông tin này là quan trọng để bảo toàn sự đa dạng trong dữ liệu.

Các bước kiểm định:

- **Bước 1:** Phân chia biến “Sleep Disorder” về thành 2 giá trị, một là không bị bệnh và 2 là bị bệnh.

- **Bước 2:** Tạo các group giữa các giá trị nhịp tim lớn hơn 78 và nhỏ hơn 78 với biến "Sleep Disorder"
- **Bước 3:** Cài đặt thư viện và kiểm định T-test (vì ở đây số mẫu lớn hơn 30)
- **Bước 4:** Kiểm tra kết quả kiểm định.

```
df['Sleep_Disorders'] = df['Sleep Disorder'].replace({'None': 1,
'Insomnia': 2, 'Sleep Apnea': 2})
```

```
import scipy.stats as stats
import pandas as pd

# Chia dữ liệu thành hai nhóm dựa trên Heart Rate
above_78 = df[df['Heart Rate'] > 78]['Sleep_Disorders']
below_78 = df[df['Heart Rate'] <= 78]['Sleep_Disorders']

# Thực hiện kiểm định t-test đối với hai nhóm
t_statistic, p_value = stats.ttest_ind(above_78, below_78,
equal_var=False)

# In kết quả
print(f'T-statistic: {t_statistic}\nP-value: {p_value}')

# Kiểm tra giả thuyết
alpha = 0.05
if p_value < alpha:
    print("Có bằng chứng để bác bỏ giả thuyết null. Có sự khác biệt
đáng kể giữa hai nhóm.")
else:
    print("Không đủ bằng chứng để bác bỏ giả thuyết null. Không có sự
khác biệt đáng kể giữa hai nhóm.")
```

Output:

```
T-Statistic: 23.463624
P-value: 1.1265e-74
Có bằng chứng để bác bỏ giả thuyết null. Có sự khác biệt đáng kể giữa
2 nhóm.
```

Như kết quả kiểm định đã cho thấy được có sự khác biệt đáng kể giữa 2 nhóm, để hiểu rõ hơn sự khác biệt như thế nào hãy xem biểu đồ dưới đây:

Có thể thấy hầu như những giá trị Heart Rate cao từ 78 đổ lên thì toàn bộ những người bị khảo sát đều bị bệnh về giấc ngủ. Vì vậy tác giả sẽ không xử lý những giá trị này.

3.3 Bộ dữ liệu đã qua tiền xử lý:

	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Heart Rate	Daily Steps	Sleep Disorder	Awakening	Blood Pressure 1	Blood Pressure 2
0	Male	27	Software Engineer	6.1	6	42	6	Overweight	77.0	4200	None	0.0	126	83
1	Male	28	Doctor	6.2	6	60	8	Normal	75.0	10000	None	0.0	125	80
2	Male	28	Doctor	6.2	6	60	8	Normal	75.0	10000	None	0.0	125	80
3	Male	28	Sales Representative	5.9	4	30	8	Obese	85.0	3000	Sleep Apnea	5.0	140	90
4	Male	28	Sales Representative	5.9	4	30	8	Obese	85.0	3000	Sleep Apnea	3.0	140	90
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
369	Female	59	Nurse	8.1	9	75	3	Overweight	68.0	7000	Sleep Apnea	3.0	140	95
370	Female	59	Nurse	8.0	9	75	3	Overweight	68.0	7000	Sleep Apnea	5.0	140	95
371	Female	59	Nurse	8.1	9	75	3	Overweight	68.0	7000	Sleep Apnea	3.0	140	95
372	Female	59	Nurse	8.1	9	75	3	Overweight	68.0	7000	Sleep Apnea	3.0	140	95
373	Female	59	Nurse	8.1	9	75	3	Overweight	68.0	7000	Sleep Apnea	0.0	140	95

374 rows × 14 columns

Bộ dữ liệu sau khi tiền xử lý bao gồm 14 cột và 374 dòng. Nhóm đã bỏ đi cột “Person ID” và tách cột “Blood Pressure” thành hai cột “Blood Pressure 1” và “Blood Pressure 2”.

```
df.isnull().sum()
missing_data_summary=pd.DataFrame({'Missing Count':
df.isnull().sum(),
                                   'Missing Percentage':
df.isnull().mean()*100})
print(missing_data_summary)
```

Output:

	Missing Count	Missing Percentage
Gender	0	0.0
Age	0	0.0
Occupation	0	0.0
Sleep Duration	0	0.0
Quality of Sleep	0	0.0
Physical Activity Level	0	0.0
Stress Level	0	0.0
BMI Category	0	0.0
Heart Rate	0	0.0
Daily Steps	0	0.0
Sleep Disorder	0	0.0
Awakening	0	0.0
Blood Pressure 1	0	0.0
Blood Pressure 2	0	0.0

Bộ dữ liệu không còn tồn tại giá trị bị thiếu.

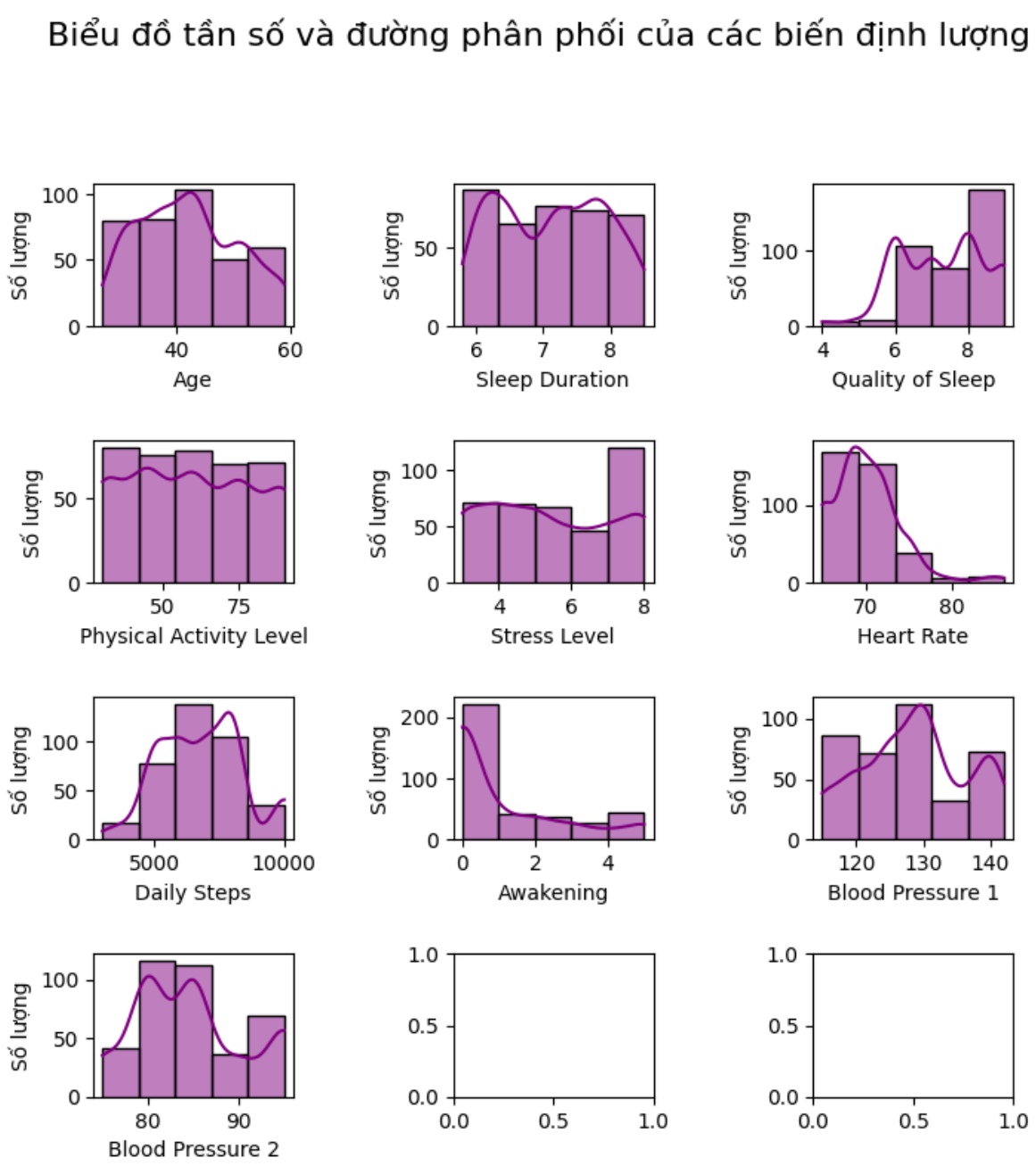
CHƯƠNG IV. PHÂN TÍCH BỘ DỮ LIỆU NGHIÊN CỨU

4.1 Phân tích đơn biến:

4.1.1 Biến định lượng:

- Thống kê mô tả của các biến định lượng:

	count	mean	std	min	25%	50%	75%	max
Age	374	42.18	8.67	27	35.25	43	50	59
Sleep Duration	374	7.13	0.8	5.8	6.4	7.2	7.8	8.5
Quality of Sleep	374	7.31	1.2	4	6	7	8	9
Physical Activity Level	374	59.17	20.83	30	45	60	75	90
Stress Level	374	5.39	1.77	3	4	5	7	8
Heart Rate	374	70.08	3.94	65	68	70	72	86
Daily Steps	374	6816.84	1617.92	3000	5600	7000	8000	10000
Awakening	374	1.09	1.61	0	0	0	2	5
Blood Pressure 1	374	128.55	7.75	115	125	130	135	142
Blood Pressure 2	374	84.65	6.16	75	80	85	90	95



- **Độ tuổi:** Độ tuổi trung bình nằm vào khoảng 42 tuổi với độ lệch là 8.67. Những người tham gia khảo sát đang ở độ tuổi trung niên.
- **Các biến về sức khỏe thể chất:**
  Các biến về sức khỏe thể chất là các biến chứa số liệu về chỉ số cơ thể và vận động của người tham gia khảo sát, bao gồm Physical Activity Level, Heart Rate, Daily Steps và Blood Pressure đã được phân thành Blood Pressure 1 (Huyết áp tâm thu) và Blood Pressure 2 (Huyết áp tâm trương).

  Phần lớn các biến trong nhóm này đều cho thấy sự phân phối khá đồng đều với độ biến động nhỏ, sự tập trung nằm trong khoảng đánh giá là bình thường về mặt sức

khỏe thể chất. Nhưng có 2 biến là Physical Activity Level và Daily Steps có sự biến động dữ liệu lớn. Cụ thể:

- Physical Activity Level (Mức độ hoạt động thể chất):
  Trung bình mức độ hoạt động thể chất (Mean): 59.17.
  Độ lệch chuẩn (Std): 20.83
- Daily Steps (Bước đi hàng ngày):
  Trung bình số bước đi hàng ngày (Mean): 6816.84.
  Độ lệch chuẩn (Std): 1617.92
  Phân phối có biên độ lớn, từ 3000 (Min) đến 10000 (Max).

→ Điều này cho thấy có sự khác biệt lớn về mặt vận động thể chất giữa những người tham gia khảo sát này.

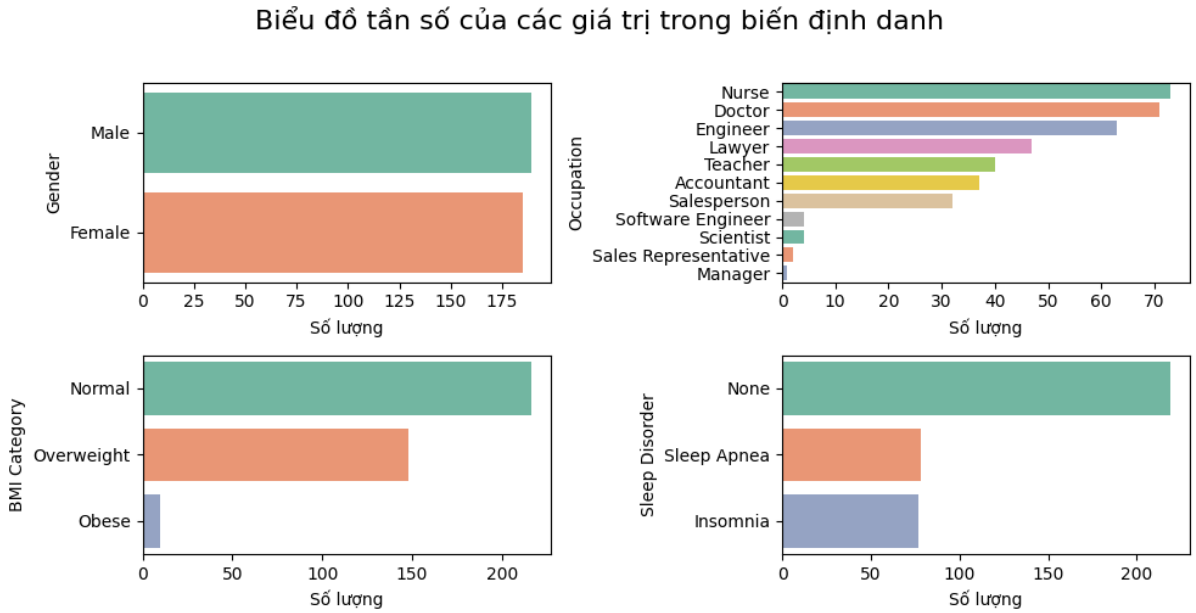
- **Biến Stress Level (Mức độ căng thẳng):**
  - Trung bình mức độ căng thẳng (Mean): 5.39.
  - Độ lệch chuẩn (Std): 1.77.
  - Mức độ căng thẳng của một người phản ánh một phần sức khỏe tinh thần của họ. Số liệu trên cho thấy những người tham gia khảo sát có mức độ căng thẳng nằm ở mức trung bình trên thang điểm 10 và từ iểu đồ phân phối của biến này, tác giả thấy đường phân phối đồng đều ở các mức điểm.

→ Có thể nói rằng trong mẫu dữ liệu này, người tham gia có mức độ căng thẳng khác nhau và trải dài trên toàn bộ thang điểm, thay vì tập trung vào một số mức độ cụ thể.

- **Các biến về chất lượng giấc ngủ:**
  - Các biến về chất lượng giấc ngủ bao gồm Sleep Duration, Quality of Sleep, Awakening thể hiện các khía cạnh liên quan đến giấc ngủ.
  - Thời gian ngủ theo lời khuyên của chuyên gia cho nhóm đối tượng trung niên rơi vào khoảng từ 7 đến 9 giờ với sai số là 1. Với số liệu thống kê bên trên thì những người tham gia khảo sát đang có thời gian ngủ khá phù hợp với độ tuổi và hầu hết đều không bị thức dậy trong đêm. Tương ứng thì họ cũng có mức độ đánh giá về chất lượng giấc ngủ hầu hết là trên mức trung bình (> 6 điểm).

→ Có thể thấy về mặt chất lượng giấc ngủ thì có những số liệu khả quan. Không có dấu hiệu cảnh báo đặc biệt về sức khỏe giấc ngủ trong nhóm tham gia.

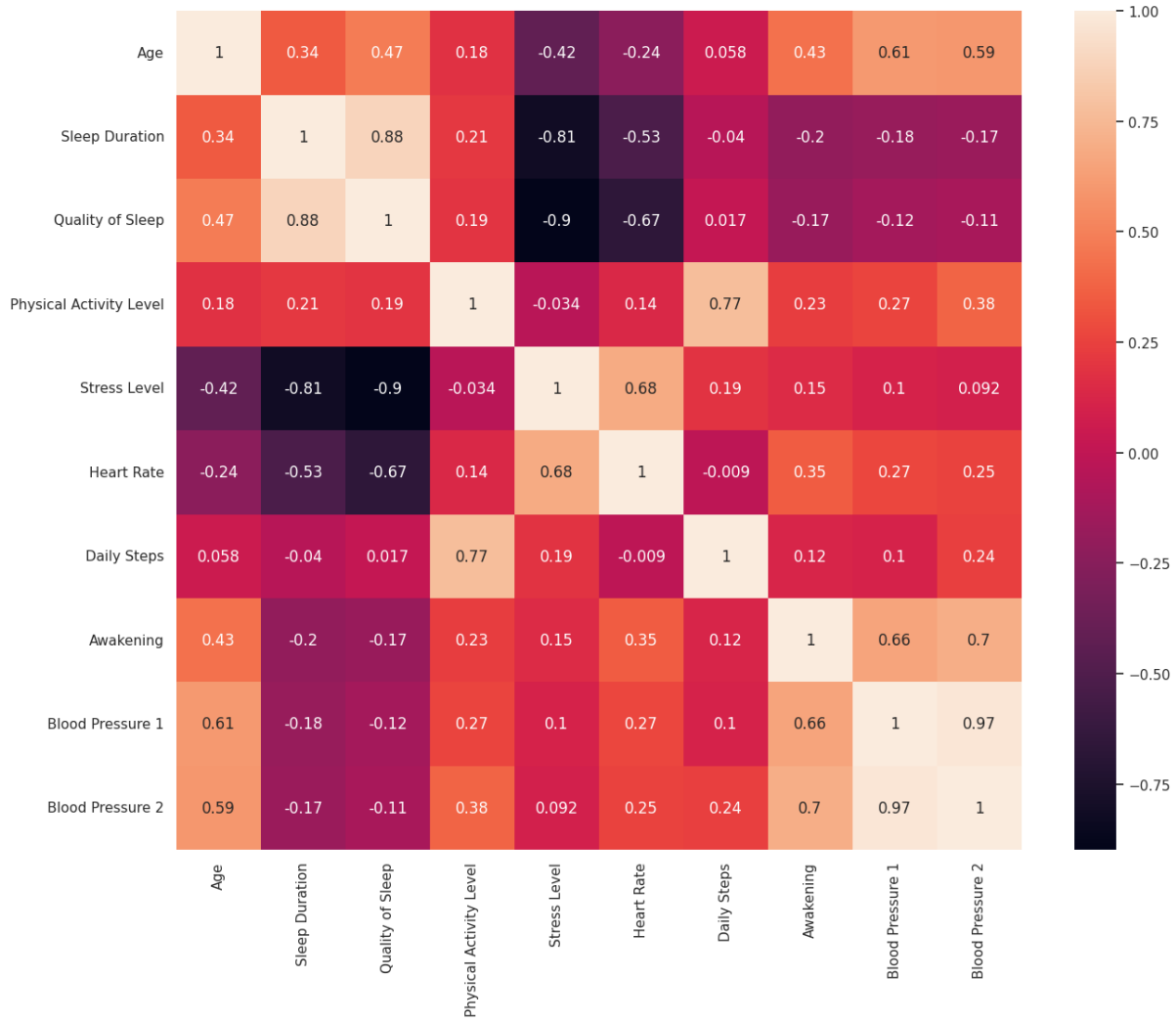
4.1.2 Biến phân loại:



- Đối tượng tham gia khảo sát đa số là nam tuy nhiên nhìn vào biểu đồ ta thấy được sự chênh lệch giữa nam và nữ là không đáng kể. Đây là một điều tốt khi từ đó ta có thể phân tích đặc điểm của mỗi giới tính mà không bị ảnh hưởng bởi chênh lệch mẫu.
- Bác sĩ và y tá là hai công việc có số lượng tham gia khảo sát nhiều nhất(hơn 70 người), ngược lại quản lý là công việc khảo sát ít nhất(khoảng 1 đến 2 người).
- Những người có chỉ số cơ thể bình thường tham gia khảo sát nhiều nhất(hơn 200 người) ngược lại những người bị bệnh béo phì (Obese) tham gia khảo sát ít nhất.(khoảng 10 người).
- Những người không bị bệnh tham gia khảo sát rất nhiều (hơn 200 người) và những người bị bệnh có số lượng tham gia tương đương nhau (khoảng 150 người).

4.2 Phân tích đa biến:

Nhóm quyết định tiến hành một số phân tích giữa 2 hay nhiều biến với nhau trong bộ dữ liệu để tìm hiểu sâu hơn về mối quan hệ giữa các biến. Mục tiêu chính là xác định các yếu tố ảnh hưởng lẫn nhau giữa các biến và tạo ra cái nhìn tổng quan về dữ liệu. Đầu tiên, nhóm sẽ đưa ra nhận xét tổng quan từ biểu đồ Heatmap về sự tương quan giữa các cặp biến.



Qua biểu đồ có thể thấy được một vài biến có sự tương quan thực sự mạnh như "Sleep Duration" và "Stress Level" hay "Quality of Sleep" cho thấy được sức khỏe tinh thần khi bị căng thẳng sẽ ảnh hưởng tới thời lượng và chất lượng giấc ngủ như thế nào. Ngoài ra cũng cho thấy được sự tương quan giữa các giá trị độ tuổi và Huyết áp, "Blood Pressure" khi có sự tương quan dương, có thể hiểu là những người càng lớn tuổi thì huyết áp sẽ càng tăng và đây cũng là một quy luật khó thể tránh khỏi.

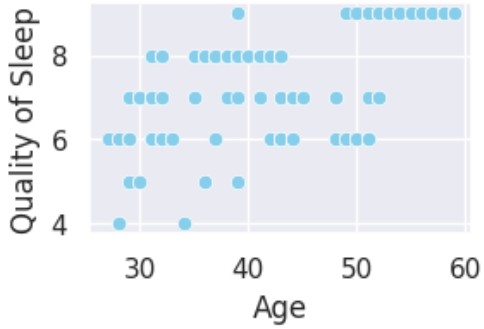
Và qua đây thì nhóm cũng đặt ra những câu hỏi như:

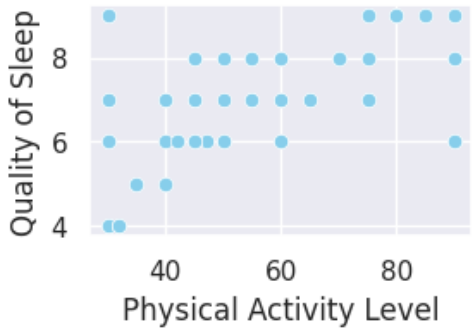
- Các ảnh hưởng của những chỉ số sức khỏe tới chất lượng giấc ngủ?
- Có sự khác biệt nào về rối loạn giấc ngủ dựa trên nghề nghiệp không?
- Sự phân bố thời gian ngủ, mức độ Stress giữa các ngành nghề, giới tính khác nhau?
- Nhịp tim của những người mắc các bệnh giấc ngủ khác nhau như thế nào?
- Có sự khác biệt về thời gian ngủ giữa các chỉ số BMI khác nhau không?
- Có mối tương quan giữa chất lượng giấc ngủ và mức độ hoạt động thể chất?
- Huyết áp của những người bị bệnh về giấc ngủ là như thế nào?
- Chúng ta có thể dự đoán sự hiện diện hay vắng mặt của chứng rối loạn giấc ngủ dựa trên các biến số đã cho không?

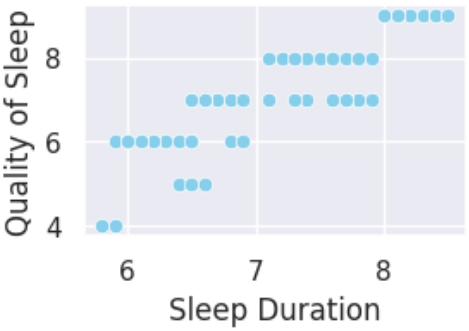
Vậy hãy cùng đi giải đáp qua phân tích đa biến và những kiểm định của nhóm sau đây.

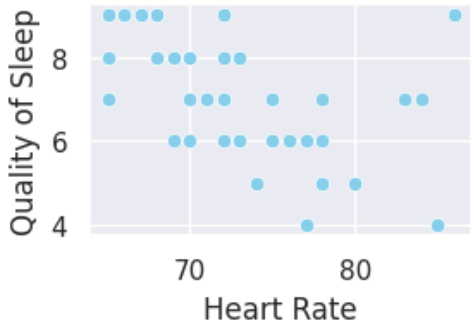
4.2.1 Chất lượng giấc ngủ

Có thể thấy được trong bộ dữ liệu, một yếu tố quan trọng để đánh giá sức khỏe giấc ngủ là "Quality of Sleep". Giá trị trong cột này được lấy trong thang điểm từ 4 đến 9, với giá trị càng cao thể hiện chất lượng giấc ngủ càng tốt. Qua việc áp dụng các phương pháp phân tích thống kê và trực quan hóa dữ liệu, nhóm hy vọng sẽ có cái nhìn rõ ràng về mối quan hệ giữa "Quality of Sleep" và các biến khác từ đó đưa ra những nhận định tổng quan về yếu các tố ảnh hưởng đến chất lượng giấc ngủ, làm nền tảng cho các quyết định và khuyến nghị trong lĩnh vực chăm sóc sức khỏe và nghiên cứu liên quan.









Qua các biểu đồ trên cho thấy được những giá trị như “Stress Level”, “Physical Activity Level” và “Sleep Duration” có sự tương quan rõ rệt với chất lượng giấc ngủ.

- Khi người khảo sát có chỉ số căng thẳng càng cao thì chất lượng giấc ngủ càng giảm.
- Khi người khảo sát dành càng nhiều thời gian cho các hoạt động thể chất, chất lượng giấc ngủ sẽ được cải thiện đáng kể.
- Khi thời lượng giấc ngủ càng cao thì chất lượng giấc ngủ từ đó cũng có sự cải thiện.

- Một số biến khác cũng cho ra những giá trị tương đối như những người dành thời gian vận động nhiều cũng như nhịp tim thấp sẽ có chất lượng giấc ngủ tốt hơn những người còn lại.

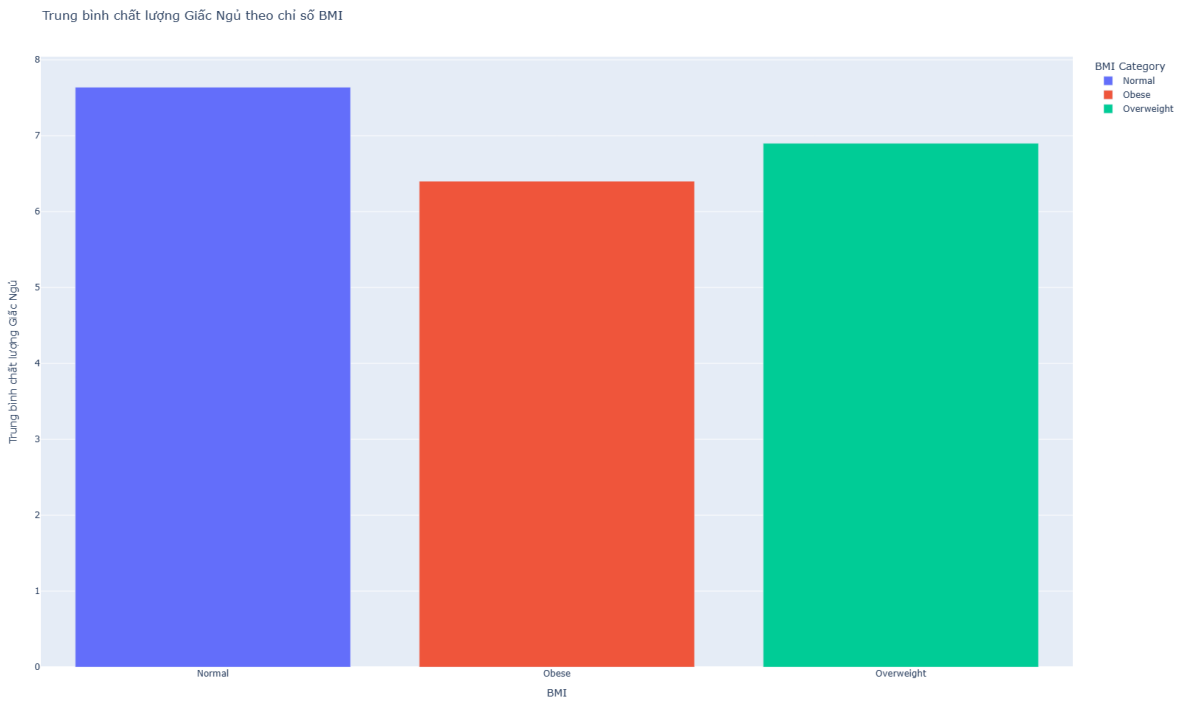
Ngoài 2 biến dễ thấy sự tuyến tính đó, như đã từng nhận xét ở trên rằng có thể có sự tuyến tính giữa biến "Quality of Sleep" và "Heart Rate". Vậy để chắc chắn hơn với những nhận định này, nhóm sẽ sử dụng kiểm định Pearson để kiểm định mối quan hệ tuyến tính giữa 2 biến trên.

```
from scipy.stats import pearsonr
X = df['Heart Rate']
Y = df['Quality of Sleep']

# Thực hiện kiểm định Pearson
corr, p_value = pearsonr(X, Y)

# In kết quả
print(f'Pearson correlation coefficient: {corr}')
print(f'P-value: {p_value}')

# Kiểm tra giả thuyết
if p_value < 0.05:
    print('Có mối tương quan tuyến tính giữa 2 biến Heart Rate và Quality of Sleep')
else:
    print('Không có mối tương quan tuyến tính giữa 2 biến Heart Rate và Quality of Sleep')
```

Output:

```
Pearson correlation coefficient: -0.6668589064754076
P-value: 1.8747152724093885e-49
Có mối tương quan tuyến tính giữa 2 biến Heart Rate và Quality of Sleep
```

Kết quả trả về cho thấy được là có mối quan hệ tuyến tính giữa nhịp tim và chất lượng giấc ngủ. Và kết quả Pearson có giá trị âm, vậy có thể nhận định được rằng khi nhịp tim càng cao thì chất lượng giấc ngủ càng giảm.



Quan sát thấy một xu hướng giảm chất lượng giấc ngủ khi chỉ số BMI thay đổi. Những người được phân loại là 'Overweight' (Thừa cân) và 'Obese' (Béo phì) cho thấy chất lượng giấc ngủ trung bình thấp hơn so với những người có chỉ số BMI 'Normal' (Bình thường).
Xu hướng này có thể là kết quả của nhiều yếu tố khác nhau. Và điều này có thể đặt ra biện pháp điều chỉnh chỉ số BMI để có thể có một sức khỏe tinh thần tốt hơn.

→ Từ những kết luận trên nhóm có thể đưa ra những nhận xét để hỗ trợ trong việc xây dựng các chiến lược chăm sóc sức khỏe và lối sống để cải thiện chất lượng giấc ngủ của người tham gia khảo sát. Khi mức độ căng thẳng hay thời lượng giấc ngủ thường bị ảnh hưởng bởi những yếu tố chủ quan và khách quan bên ngoài, thì tăng cường thời gian tập thể dục là một giải pháp đáng cân nhắc, việc tập thể dục sẽ giúp giấc ngủ sâu hơn và dài hơn và từ đó tạo tiền đề để cải thiện chất lượng giấc ngủ của chính bản thân mình.

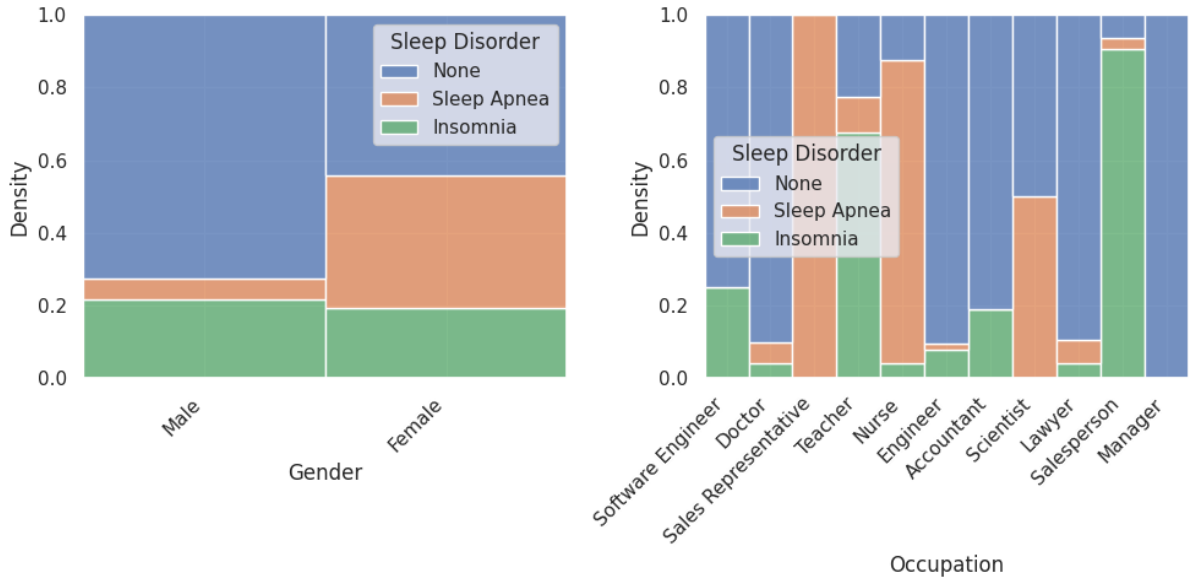
4.2.2 Bệnh về giấc ngủ:

a. Giới tính và ngành nghề:

Với một trong các mục đích chính của bộ dữ liệu là xác định và dự báo được những cá nhân bị bệnh về giấc ngủ, thì tầm ảnh hưởng của yếu tố giới tính và nghề nghiệp cũng là một vấn đề quan trọng, đặc biệt là khi có sự khác biệt sinh học giữa nam và nữ cũng như đặc điểm đặc trưng của các ngành nghề.

```
import warnings
warnings.filterwarnings("ignore")
cat_vars = ['Gender', 'Occupation', 'BMI Category']
fig, axs = plt.subplots(nrows=1, ncols=3, figsize=(15, 5))
axs = axs.flatten()
```

```
for i, var in enumerate(cat_vars):
    sns.histplot(x=var, hue='Sleep Disorder', data=df, ax=axs[i],
    multiple="fill", kde=False, element="bars", fill=True,
stat='density')
    axs[i].set_xticklabels(df[var].unique(), rotation=45, ha='right')
    axs[i].set_xlabel(var)
fig.tight_layout()
plt.show()
```



Qua biểu đồ tác giả có thể thấy được:

- Giới tính và Bệnh về giấc ngủ:
    - Tỉ lệ mắc hội chứng mất ngủ ở nam và nữ là như nhau ở bộ dữ liệu này, tuy nhiên trên tổng quan thì tỉ lệ nữ giới mắc bệnh liên quan tới giấc ngủ cao hơn khá nhiều so với nam giới khi chiếm tới gần 50%.
    - Các nghiên cứu trước đây đã chỉ ra rằng nữ có thể trải qua các vấn đề giấc ngủ khác biệt so với nam, chẳng hạn như chu kỳ kinh nguyệt và thai kỳ. Đây cũng là lý do tỷ lệ mắc bệnh về giấc ngủ cao hơn nam giới.

Tuy nhiên tỷ lệ này cũng bị ảnh hưởng bởi tập dữ liệu đầu vào, và hãy xem tiếp về nghề nghiệp để hiểu rõ hơn vấn đề này:

- Nghề nghiệp và Bệnh về giấc ngủ:
    - Mỗi ngành nghề có thể mang đến những đặc tính đặc biệt, ảnh hưởng đến chế độ làm việc và áp lực công việc, có thể gây ảnh hưởng đến giấc ngủ của nhân viên.
    - Như ở trong bộ dữ liệu này có thể thấy được những ngành nghề có tỉ lệ mắc bệnh về giấc ngủ cao như:
        - ➔ Insomnia: Teacher, Salesperson, Software Engineer, Accountant.

➔ Sleep Apnea: Sales Representative, Nurse và Scientist.

Qua quan sát cho thấy được những ngành nghề bị bệnh Insomnia cao không có sự khác biệt về tỉ lệ giới tính. Vì vậy tỉ lệ bị bệnh Insomnia ở bộ dữ liệu này giữa nam và nữ như nhau là điều dễ hiểu. Tiếp theo về Sleep Apnea thì có sự chênh lệch khi chủ yếu các y tá đều là nữ. Từ đó dẫn đến tỷ lệ bệnh Sleep Apnea của nữ giới cao hơn hẳn với nam giới ở bộ dữ liệu này.
→ Thông qua việc khám phá tác động của giới tính và nghề nghiệp, nhóm đã đạt được cái nhìn toàn diện hơn về những yếu tố ảnh hưởng đến giấc ngủ và cung cấp cơ sở dữ liệu hữu ích cho việc xây dựng mô hình dự báo bệnh giấc ngủ cũng như giúp nhận định được những ngành nghề nào dễ bị bệnh về giấc ngủ hơn những ngành khác.

b. Mức độ căng thẳng:

Ngoài ra, khi đã phân tích các yếu tố ngành nghề ảnh hưởng tới bệnh về giấc ngủ thì không thể bỏ qua mức độ căng thẳng, khi ngành nghề công việc ảnh hưởng trực tiếp tới vấn đề này. Và để chắc chắn hơn về nhận định này, nhóm quyết định kiểm định với mức độ tin cậy 95% xem rằng "Stress Level" có thực sự ảnh hưởng tới "Sleep Disorder" hay không. Các bước sẽ như sau:

- Đầu tiên nhóm định dạng lại biến "Sleep Disorder" để thuận tiện cho kiểm định, cũng như sẽ chia mức độ căng thẳng thành 3 cấp độ là thấp, trung bình, cao.
- Tiếp theo sẽ là biểu diễn Crosstab:

```
df['Stress_Level'] = df['Stress Level'].replace({
    3.0: 'Thấp',
    4.0: 'Thấp',
    5.0: 'Trung bình',
    6.0: 'Trung bình',
    7.0: 'Cao',
    8.0: 'Cao'})
Marital=df[['Sleep Disorder','Stress_Level']]
crosstab = pd.crosstab(Marital["Sleep Disorder"],
Marital["Stress_Level"])
crosstab
```

Output:

Stress_Level	Cao	Thấp	Trung bình
Sleep Disorder			
Insomnia	44	25	8
None	36	83	100
Sleep Apnea	40	33	5

- Và chạy kiểm định:

```
contingency_table = pd.crosstab(df['Stress Level'], df['Sleep
Disorder'])
chi2_stat, p_value, _, _ = chi2_contingency(contingency_table)

print(f"Chi-squared statistic: {chi2_stat}")
print(f"P-value: {p_value}")

alpha = 0.05
if p_value < alpha:
    print("Có đủ bằng chứng để bác bỏ H0. Có sự tương quan giữa 2
biến")
else:
    print("Không đủ bằng chứng để bác bỏ H0. Không có sự tương quan
giữa 2 biến")
```

Output:

```
Chi-squared statistic: 240.19936847934278
P-value: 6.221717380449499e-46
Có đủ bằng chứng để bác bỏ H0. Có sự tương quan giữa 2 biến
```

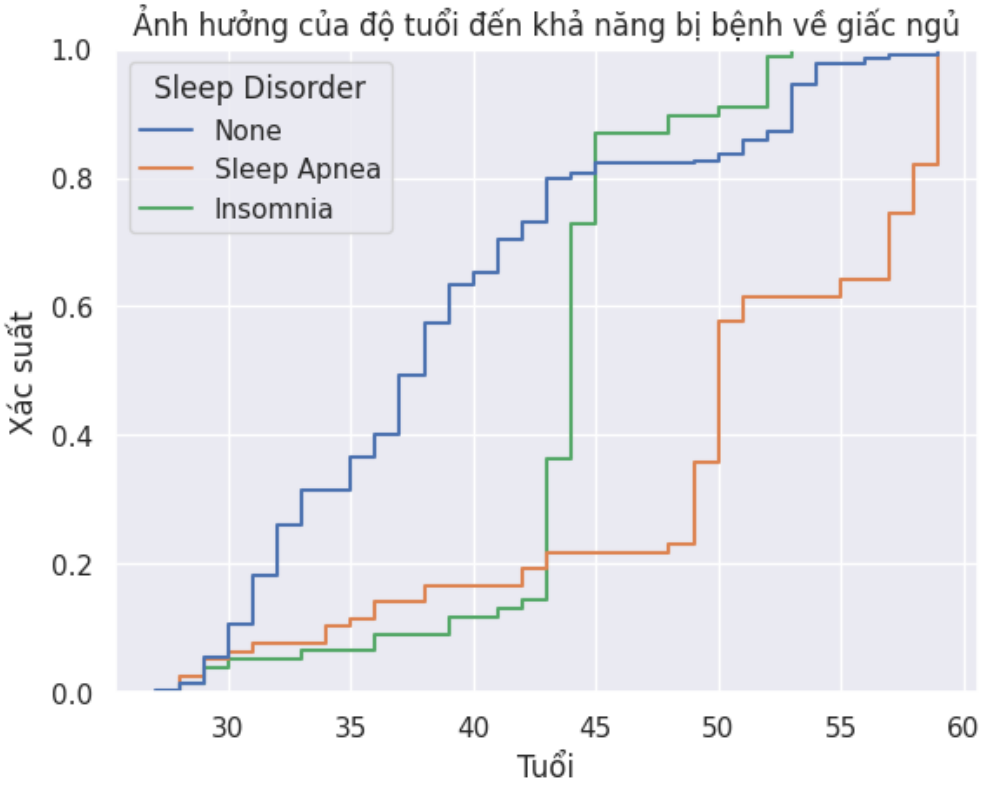
Nhận xét: Kết quả trả về tỉ lệ P-value là rất thấp và bác bỏ H0, từ đó cho thấy được mức độ căng thẳng thực sự ảnh hưởng tới bệnh về giấc ngủ. Qua đây có thể đưa ra được những kết luận sơ bộ rằng những người có mức độ căng thẳng cao sẽ dẫn tới mức độ sức khỏe giấc ngủ giảm đáng kể và tăng tỉ lệ mắc các bệnh liên quan tới giấc ngủ. Và thông qua thông tin này sẽ giúp những người bệnh có thêm vấn đề đáng cân nhắc để có thể cải thiện tình trạng giấc ngủ của mình.

c. Độ tuổi:

Theo các thống kê nghiên cứu trên thế giới đã chỉ ra rằng người già thường gặp vấn đề về giấc ngủ. Có thể là do sức khỏe yếu đi khi lớn tuổi cũng như các sự thay đổi về cấu trúc giấc ngủ cũng làm góp phần làm tăng nguy cơ mắc các vấn đề giấc ngủ ở người già. Vậy hãy thử xem trong bộ dữ liệu này thì tỉ lệ mắc các bệnh về giấc ngủ theo độ tuổi bằng bản đồ ecdf sẽ như thế nào:

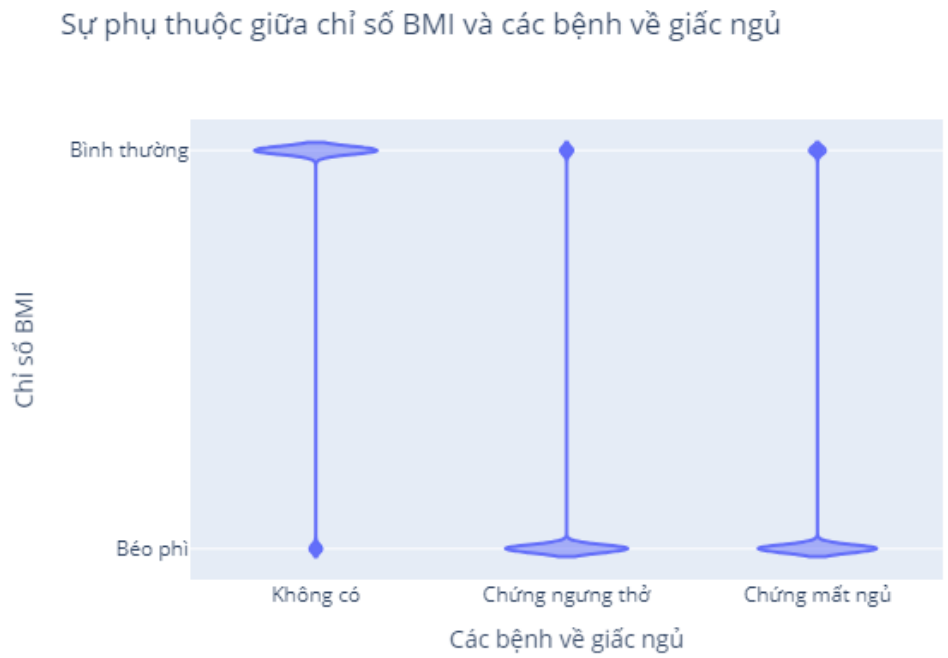
```
sns.ecdfplot(data=df, x='Age', hue='Sleep Disorder')
plt.xlabel('Tuổi')
plt.ylabel('Xác suất')
plt.title('Ảnh hưởng của độ tuổi đến khả năng bị bệnh về giấc ngủ',
```

```
fontsize=12)
plt.show()
```



Nhận xét: Qua biểu đồ tác giả thấy được rằng từ mốc 40 tuổi trở đi, khả năng bị bệnh Insomnia và Sleep Apnea tăng lên đáng kể, qua đó cho thấy được những người cao tuổi có khả năng bị bệnh về giấc ngủ cao hơn so với nhóm người trẻ hơn. Để chắc chắn về vấn đề này thì nhóm sẽ kiểm định ở phần sau.

d. Chỉ số BMI:



Qua biểu đồ tác giả có thể thấy được rằng số lượng người bị bệnh về giấc ngủ khi có tình trạng BMI là "Overweight" và "Obese" là rất cao so với số lượng người trong mỗi nhóm.

Nhận xét: Từ đây có thể nhận thấy được chỉ số BMI cũng có thể là tác nhân tác động tới việc người bị bệnh về giấc ngủ hay không. Từ đây có thể đưa ra được một trong những biện pháp cải thiện sức khỏe giấc ngủ là giữ cho mình một chỉ số BMI vừa phải thông qua việc sinh hoạt điều độ và tập luyện thường xuyên.

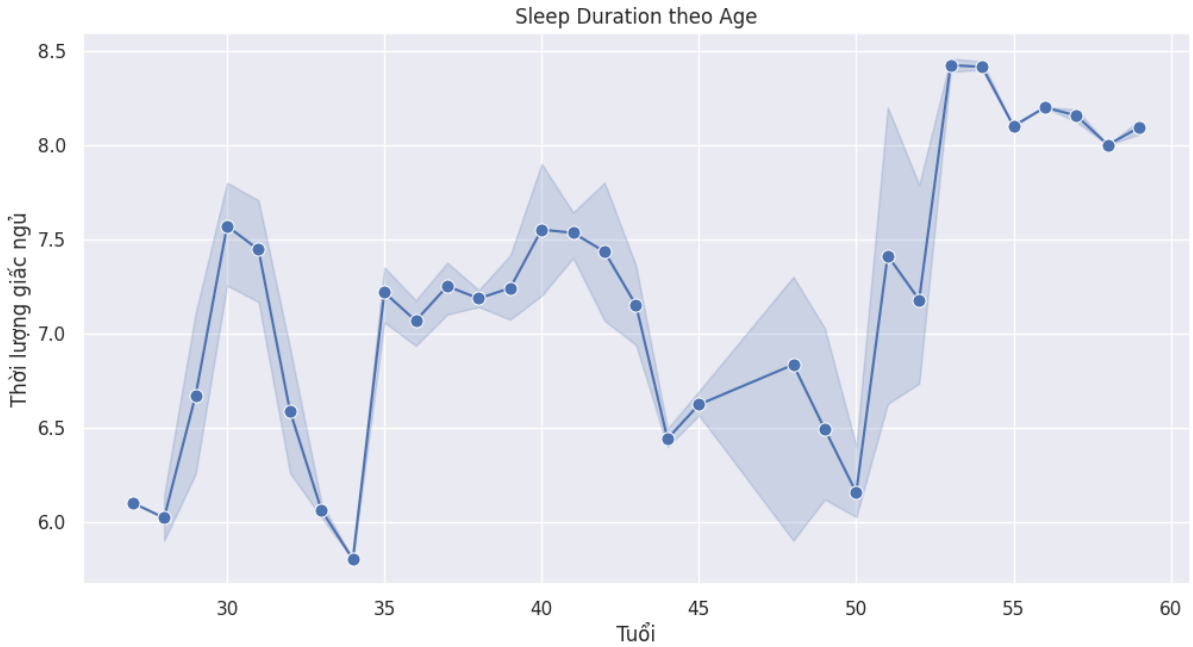
- ❖ **Tổng kết:**

Nhìn chung, qua việc phân tích các yếu tố như giới tính, nghề nghiệp, mức độ căng thẳng, độ tuổi, và chỉ số BMI, tác giả nhận thức được rằng tình trạng bệnh về giấc ngủ không chỉ phụ thuộc vào yếu tố giới tính mà còn chịu ảnh hưởng sâu sắc từ môi trường làm việc và lối sống. Đối mặt với những thách thức này, việc xây dựng một chiến lược chăm sóc sức khỏe giấc ngủ toàn diện trở nên quan trọng hơn bao giờ hết để cải thiện chất lượng cuộc sống.

Tuy nhiên, những nhận định này chỉ là mức đánh giá theo các biểu đồ. Nhóm sẽ tiếp tục kiểm định lại các vấn đề này trong chương tiếp theo, nhằm đảm bảo rằng kết quả đưa ra sẽ là chính xác nhất dựa trên bộ dữ liệu. Điều này sẽ giúp tác giả có góc nhìn toàn diện và chuẩn xác hơn, cung cấp thông tin chất lượng để hỗ trợ quá trình xây dựng chiến lược chăm sóc sức khỏe giấc ngủ.

4.2.3 Các yếu tố sức khỏe khác:

a. Độ tuổi là một vấn đề thường được nhắc tới đầu tiên khi nói về sức khỏe. Vậy hãy xem qua một vài yếu tố có thể liên quan tới tuổi tác trong bộ dữ liệu này:

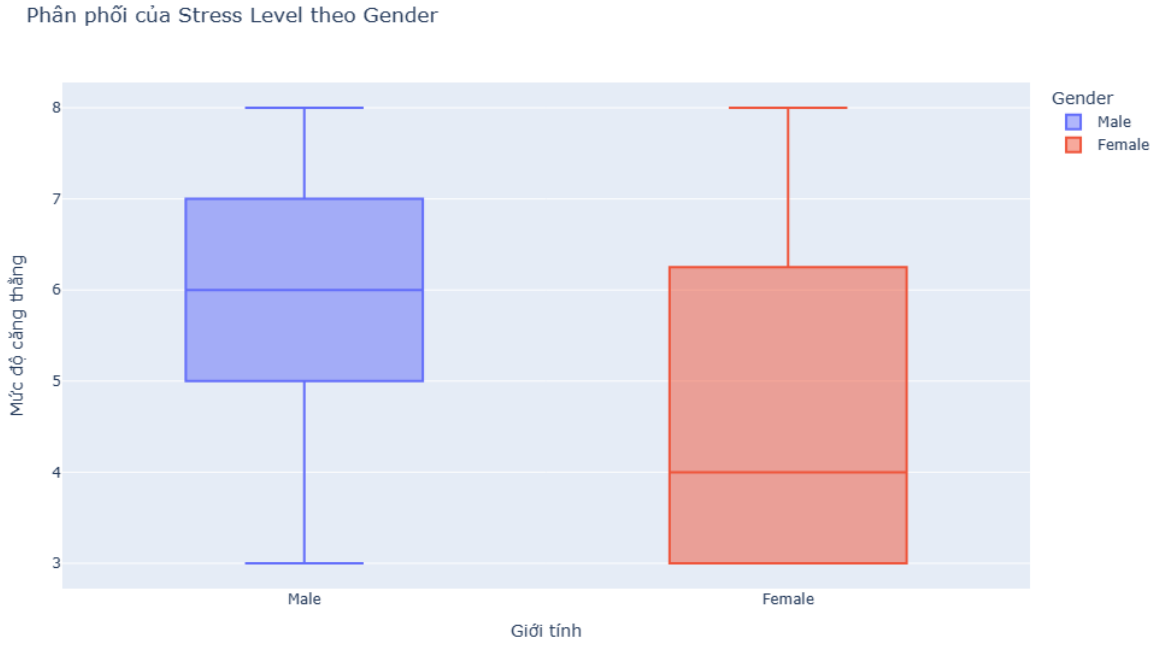


Qua biểu đồ trên tác giả có thể đưa ra nhận xét rằng:

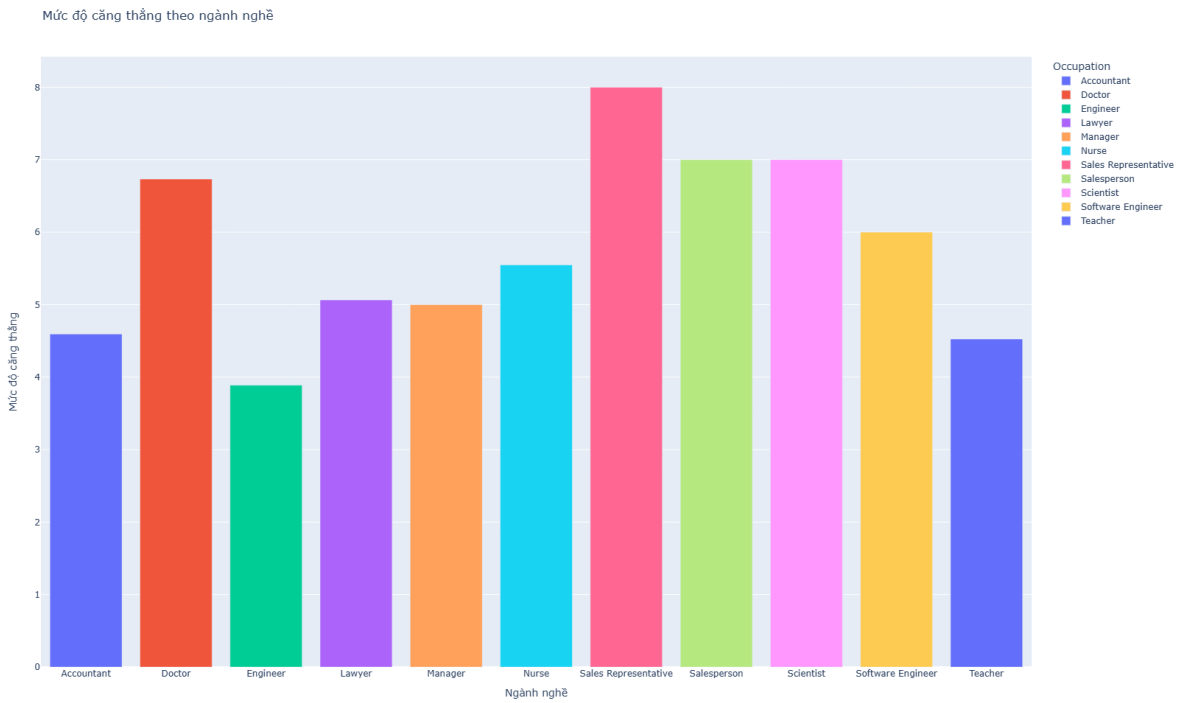
- Thời lượng giấc ngủ và mức độ căng thẳng theo độ tuổi:
    - Nhóm người trên 50 tuổi có xu hướng có thời lượng giấc ngủ cao hơn so với các nhóm tuổi khác. Có thể giải thích điều này bằng việc những người trên 50 tuổi thường đã tới độ tuổi nghỉ hưu, không cần phải đối mặt với áp lực công việc hàng ngày nữa. Điều này có thể giúp họ có thêm thời gian cho giấc ngủ và duy trì một thời lượng giấc ngủ tốt hơn.
    - Ngoài ra, sự giảm áp lực từ công việc có thể giúp giảm stress và lo âu, hai yếu tố thường xuyên ảnh hưởng đến chất lượng giấc ngủ. Do đó, nhóm người trên 50 tuổi có thể tận hưởng một giấc ngủ sâu hơn và hơn nữa, có thể dành thời gian cho việc điều chỉnh và duy trì thói quen ngủ lành mạnh.

→ Tuy nhiên, cần lưu ý rằng có sự biến động lớn trong mẫu dữ liệu và có những yếu tố khác nhau có thể ảnh hưởng đến thời lượng giấc ngủ, bao gồm cả yếu tố về sức khỏe, môi trường sống, và lối sống cá nhân.

b. Tiếp theo là mức độ căng thẳng khi như những mối tương quan đã thấy ở trên thì mức độ căng thẳng ảnh hưởng trực tiếp tới sức khỏe giấc ngủ của những người được khảo sát.



Như biểu đồ đã thể hiện được mức độ căng thẳng trung bình của nam giới cao hơn so với nữ giới. Kể cả qua mức độ phân phối hay là đường trung bình được biểu diễn. Điều này cũng có thể dự đoán trước được khi thời lượng ngủ và chất lượng giấc ngủ của nam giới cũng thấp hơn. Và qua đây chó thấy được mối liên kết giữa các chỉ số tinh thần, thời lượng ảnh hưởng như thế nào tới sức khỏe giấc ngủ của mỗi người.



Như phần trước đã phân tích về ảnh hưởng của nghề nghiệp với tình trạng bệnh về giấc ngủ, còn mức độ căng thẳng thì như thế nào. Khi xem xét mối quan hệ giữa nghề nghiệp và mức độ căng thẳng, có vẻ như 'Sale Representative' và 'Scientist' thể hiện mức độ căng thẳng cao nhất. Tuy nhiên, đây chỉ là trong tập dữ liệu mẫu này. Mức độ căng thẳng liên quan đến nghề nghiệp có thể rất khác nhau tùy thuộc các yếu tố khách quan và chủ quan khác. Các yếu tố như nhu cầu công việc, môi trường làm việc, văn hóa công ty và bản thân mỗi cá nhân đều có thể ảnh hưởng đến mức độ căng thẳng của bản thân.

Do đó, những kết quả trên được biểu diễn ra để nhóm hiểu hơn về bộ dữ liệu và có cái nhìn tổng quan hơn về sự ảnh hưởng của mức độ căng thăng lên các ngành nghề. Từ đó các đại diện doanh nghiệp có thể nhìn vào số liệu thống kê và đưa ra những biện pháp xử lý giúp nhân viên tránh khỏi sự căng thẳng trong công việc.

Chương V: KIỂM ĐỊNH

Ở phần này, tác giả sẽ đi kiểm định những giả thuyết được đưa ra để có thể hiểu hơn về bộ dữ liệu, cũng như làm rõ những nhận định chưa chắc chắn ở các phần trên.

5.1 Các yếu tố ảnh hưởng tới thời lượng và chất lượng giấc ngủ

5.1.1 Giới tính

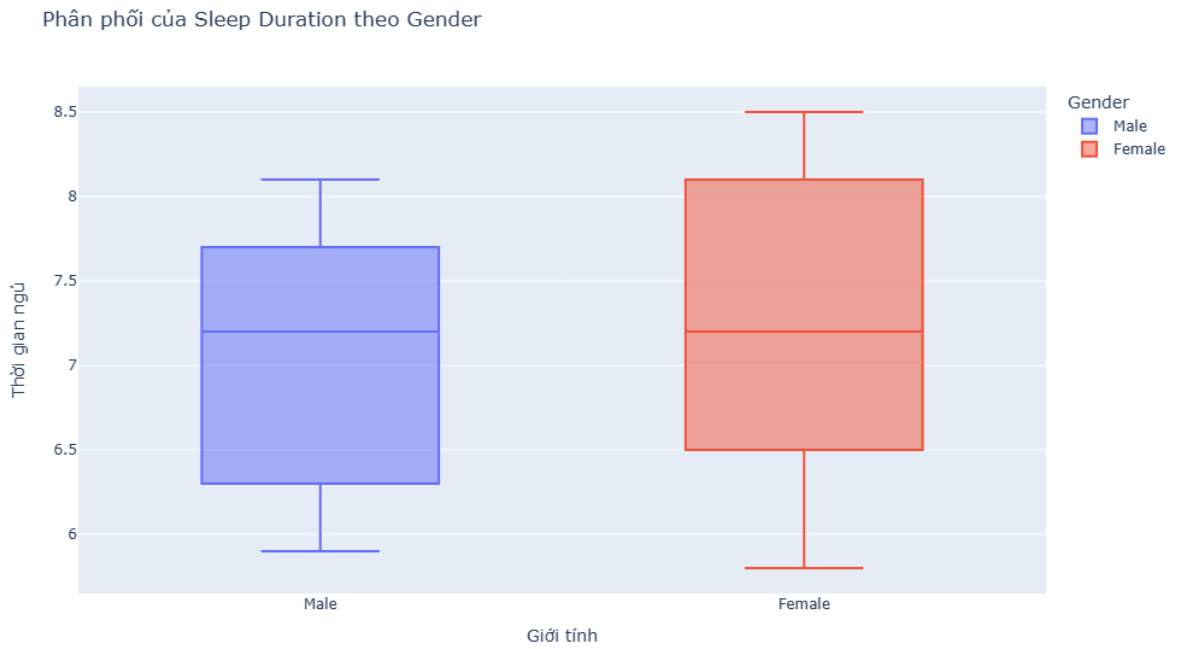
Đầu tiên yếu tố giới tính là một yếu tố cần phải xem xét qua khi số lượng nữ giới và nam giới trong bộ dữ liệu này là không có sự khác biệt. Đây là điều kiện hợp lý để xem xét những mối liên quan của giới tính tới các yếu tố khác trong bộ dữ liệu:

```
df['Gender'].value_counts()
```

Output:

```
Male      189
Female    185
Name: Gender, dtype: int64
```

a. Mối tương quan giữa nam và nữ trong thời lượng giấc ngủ:



Qua biểu đồ boxplot thì tác giả có thể đưa ra nhận định sơ bộ được rằng trong bộ dữ liệu này, nữ giới có thời lượng ngủ nhiều hơn nam giới, khi miền giá trị của nữ giới kéo dài tới 8.5. Tuy nhiên để xác định điều này là đúng, hãy kiểm định với mức độ tin cậy là 95%.

- **Kiểm định giả thuyết:** Nam và nữ có thời lượng giấc ngủ là như nhau:

  $H0$: $\mu\{Sleep\ Duration\}[Male] = \mu\{Sleep\ Duration\}[Female]$

$H1$: $\mu\{Sleep\ Duration\}[Male] \neq \mu\{Sleep\ Duration\}[Female]$

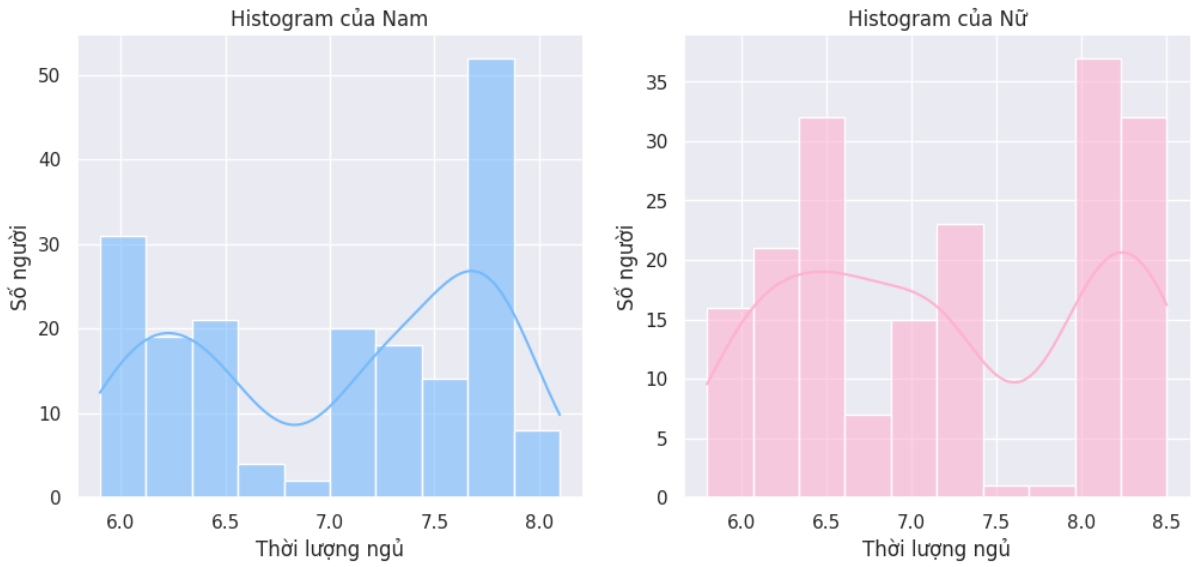
```
df2 = df[['Gender', 'Sleep Duration']]
male_group = df2[df2["Gender"] == "Male"]["Sleep Duration"].dropna()
female_group = df2[df2["Gender"] == "Female"]["Sleep
Duration"].dropna()
# Thực hiện kiểm định t-test
t_statistic, p_value = stats.ttest_ind(male_group, female_group,
equal_var=False)
print("T-statistic:", t_statistic)
print("Giá trị p:", p_value)
alpha = 0.05
if p_value < alpha:
    print("Có đủ bằng chứng để bác bỏ H0.")
else:
    print("Không đủ bằng chứng để bác bỏ H0.")
```

Output:

```
T-statistic: -2.356536840262248
Giá trị p: 0.018997711592424273
Có đủ bằng chứng để bác bỏ H0.
```

Ở đây kết quả trả về là bác bỏ H0, chứng tỏ rằng Nam và Nữ có sự khác biệt đáng kể về thời lượng giấc ngủ.

- Biểu đồ thể hiện:



Nhận xét: Qua biểu đồ thì tác giả thấy rõ hơn được về phân phối của thời lượng giấc ngủ mà

nam giới và nữ giới có sự chênh lệch và cụ thể là nữ giới trong tập dữ liệu có thời lượng ngủ nhiều hơn nam giới khi những cột mốc chủ yếu ở mức 6.5, 8.0 và 8.5 giờ trong khi ở nam giới chỉ là 6.0, 7.0 và 7.8 giờ.

b. Mối tương quan giữa nam và nữ trong chất lượng giấc ngủ:

Tương tự như ở trên, nhóm đặt ra giả thuyết:

- Kiểm định giả thuyết: Nam và nữ có chất lượng giấc ngủ là như nhau:

  $H0$: $\mu\{Quality\ of\ Sleep\}[Male] = \mu\{Quality\ of\ Sleep\}[Female]$

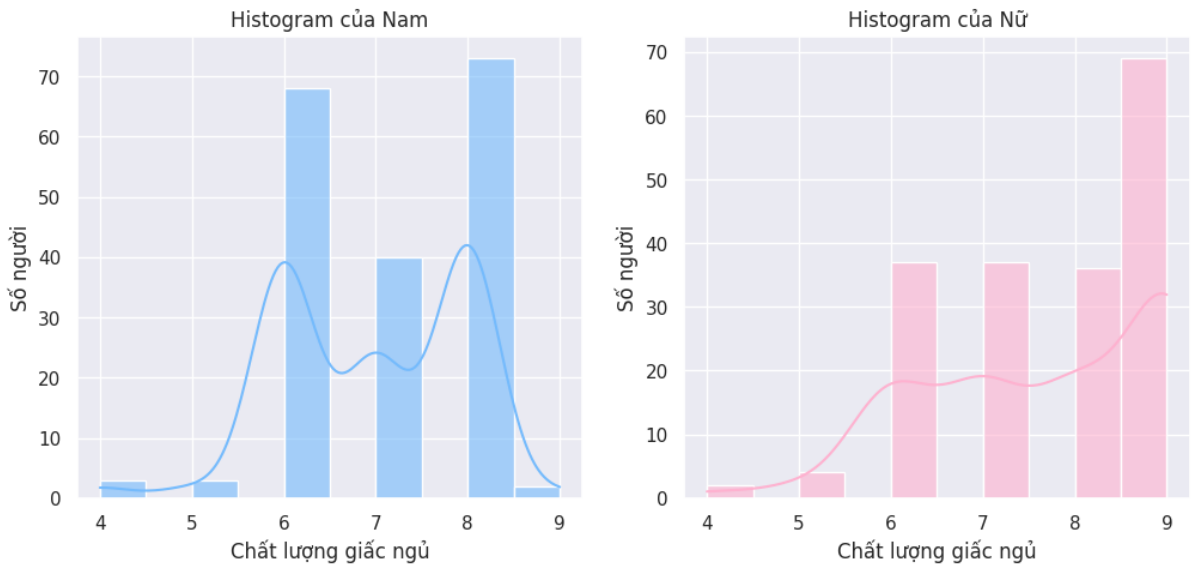
  $H1$: $\mu\{Quality\ of\ Sleep\}[Male] \neq \mu\{Quality\ of\ Sleep\}[Female]$

```
df2 = df[['Gender', 'Quality of Sleep']]
male_group = df2[df2["Gender"] == "Male"]["Quality of
Sleep"].dropna()
female_group = df2[df2["Gender"] == "Female"]["Quality of
Sleep"].dropna()
# Thực hiện kiểm định t-test
t_statistic, p_value = stats.ttest_ind(male_group, female_group,
equal_var=False)
print("T-statistic:", t_statistic)
print("Giá trị p:", p_value)
alpha = 0.05
if p_value < alpha:
    print("Có đủ bằng chứng để bác bỏ H0.")
else:
    print("Không đủ bằng chứng để bác bỏ H0.")
```

Output:

```
T-statistic: -5.859317976055214
Giá trị p: 1.078122839334259e-08
Có đủ bằng chứng để bác bỏ H0.
```

- Biểu đồ thể hiện:



Nhận xét:

- Như đã được trình bày trước đó, tác giả có kết luận được rằng có sự tương quan dương giữa chất lượng giấc ngủ và thời gian ngủ. Và thông qua 2 kiểm định trên, tác giả càng chắc chắn hơn về nhận đình này.
- Trước đó, nữ giới có thời lượng ngủ cao hơn nam giới. Và bây giờ ở đây thì ta thấy được chất lượng giấc ngủ của nữ giới duy trì ở mức trung bình đổ lên, đặc biệt là rất ổn định ở các mức điểm 6,7,8 và một số lượng lớn là 9. Ngược lại, chất lượng giấc ngủ của nam giới thường tập trung ở mức 6 và 8 điểm và rất ít số người đạt tới 9 điểm.

Qua đây càng cho thấy được có vẻ như nữ giới có tổng quan giấc ngủ tốt hơn nam giới và đánh giá lại được quan điểm về sự tương quan giữa "Quality of Sleep" và "Sleep Duration" thông qua các biểu đồ và kiểm định với biến giới tính.

5.1.2 Huyết áp và thời lượng giấc ngủ

Các nghiên cứu cho thấy những người ngủ ít thường dẫn tới việc huyết áp cao hơn những người bình thường. Vậy hãy cùng thử nghiệm kiểm định giả thuyết đó trong bộ dữ liệu này:
Giả thuyết: Huyết áp tâm thu không có mối tương quan với thời lượng giấc ngủ.
Phương pháp kiểm định: Kiểm định Pearson.

- Cách thức đánh giá:
    - Nếu giá trị tuyệt đối của r < 0.1, thì với kết quả nhỏ như vậy, nhóm xem xét là không hề có mối quan hệ tương quan tuyến tính đáng kể ở đây.
    - Nếu 0.1 <= |r| < 0.3, mối quan hệ được xem là rất yếu.
    - Nếu 0.3 <= |r| < 0.5, mối quan hệ được xem là trung bình.
    - Nếu |r| >= 0.5, mối quan hệ được xem là mạnh.

(Những chỉ số đánh giá này dựa theo ý kiến chủ quan của nhóm nghiên cứu cũng như ngữ cảnh của bộ dữ liệu)

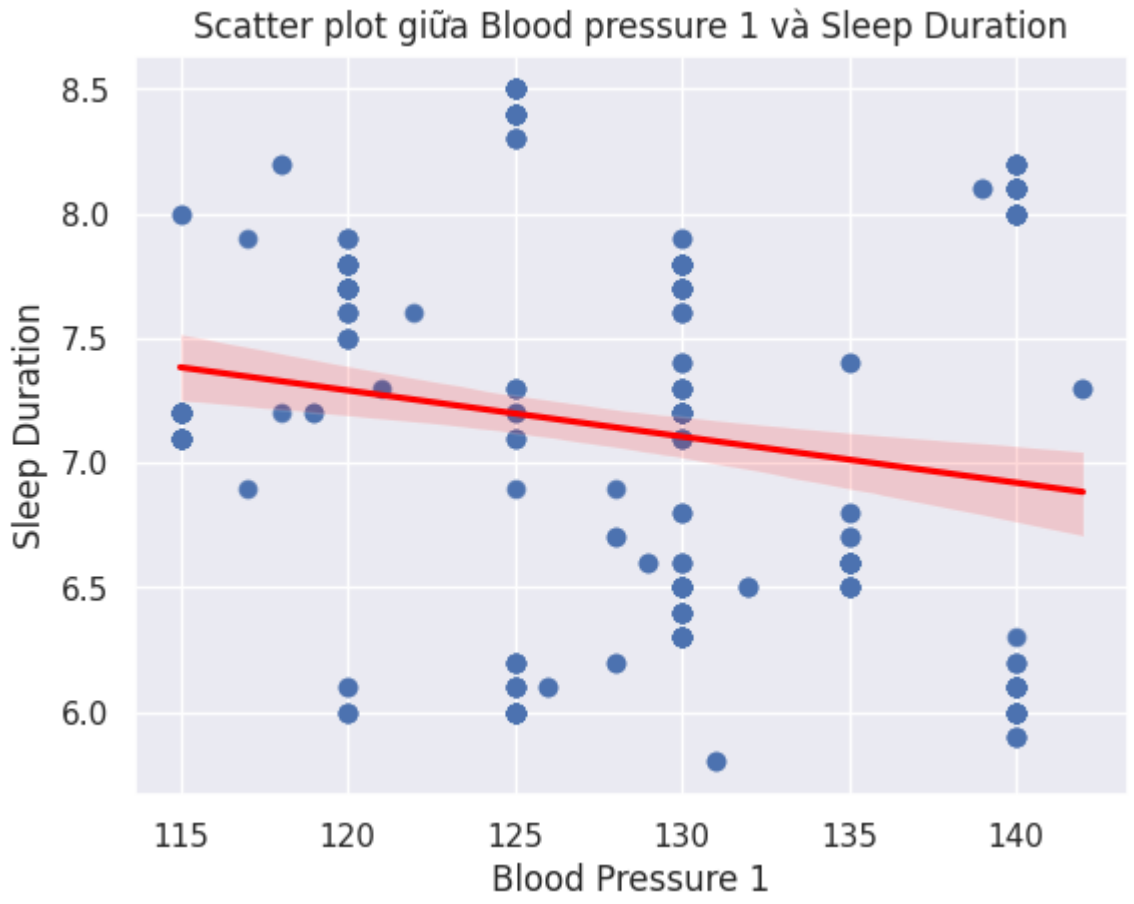
```
# Tính hệ số tương quan Pearson và p-value
r, p_value = pearsonr(df['Blood Pressure 1'], df['Sleep Duration'])
r = round(r, 2)
p_value = round(p_value, 4)

print(f"Pearson correlation coefficient (r): {r}")
print(f"P-value: {p_value}")
if np.sign(r) == -1:
    s = "âm"
elif np.sign(r) == 1:
    s = "dương"
if np.abs(r) < 0.1:
    print('Huyết áp tâm thu "Blood Pressure 1" và thời lượng giấc ngủ
không có mối quan hệ tương quan tuyến tính')
elif 0.1 <= np.abs(r) < 0.3:
    print('Huyết áp tâm thu "Blood Pressure 1" và thời lượng giấc ngủ
có mối tương quan tuyến tính', s, 'rất yếu')
elif 0.3 <= np.abs(r) < 0.5:
    print('Huyết áp tâm thu "Blood Pressure 1" và thời lượng giấc ngủ
có mối tương quan tuyến tính', s, 'trung bình')
else:
    print('Huyết áp tâm thu "Blood Pressure 1" và thời lượng giấc ngủ
có mối tương quan tuyến tính', s, 'mạnh')
```

Output:

```
Pearson correlation coefficient (r): -0.18
P-value: 0.0005
Huyết áp tâm thu "Blood Pressure 1" và thời lượng giấc ngủ có mối
tương quan tuyến tính âm rất yếu
```

Nhận xét: Qua kết quả kiểm định cho thấy được:

- Pearson correlation coefficient (r): Hệ số tương quan Pearson là -0.18. Giá trị này thể hiện một mối quan hệ tương quan âm giữa "Blood Pressure 1" và "Quality of Sleep." Điều này cho thấy rằng khi thời lượng giấc ngủ giảm xuống thì huyết áp tâm thu sẽ có xu hướng tăng lên một lượng nhỏ và điều này phù hợp với nhận định ban đầu của nhóm.



Biểu đồ biểu diễn cho thấy được rõ hơn các nhận định trên.

- Mặc dù còn một vài điểm ngoài vùng làm ảnh hưởng kết quả tương quan những nhìn trên tổng thể những giá trị huyết áp thấp xuất hiện ở thời lượng giấc ngủ từ 7 giờ đổ lên và những giá trị huyết áp cao từ 125 tới 135 chủ yếu ở mức thời lượng ngủ dưới 7.5 giờ.

Vậy kiểm định này cho thấy được một mối quan hệ tương quan nghịch giữa thời lượng giấc ngủ và huyết áp của người được khảo sát. Ngoài ra, kết luận của kiểm định cũng phù hợp với nhận định ban đầu. Qua đây cho thấy được việc tăng thời lượng ngủ sẽ giúp cải thiện một phần nào đó huyết áp, cũng như kéo theo chất lượng giấc ngủ cao hơn.

5.2 Các yếu tố ảnh hưởng tới bệnh về giấc ngủ

5.2.1 Giới tính và bệnh về giấc ngủ

Như ở phía trên nhóm đã đưa ra nhận định rằng nữ giới mắc bệnh về giấc ngủ nhiều hơn nam giới trong bộ dữ liệu này. Vậy hãy thử kiểm định lại xem nhận định trên có đúng hay không.

Giả thuyết: Tỷ lệ bị bệnh về giấc ngủ của nam và nữ là như nhau:

$$H0:\ \mu\{Sleep\ Disorder\}[Male] = \mu\{Sleep\ Disorder\}[Female]$$

$$H1:\ \mu\{Sleep\ Disorder\}[Male] \neq \mu\{Sleep\ Disorder\}[Female]$$

```
Marital=df[['Sleep Disorder','Gender']]
crosstab = pd.crosstab(Marital["Sleep Disorder"], Marital["Gender"])
crosstab
```

Output:

Gender	Female	Male
Sleep Disorder		
Insomnia	36	41
None	82	137
Sleep Apnea	67	11

```
contingency_table = pd.crosstab(df['Gender'], df['Sleep Disorder'])
chi2_stat, p_value, _, _ = chi2_contingency(contingency_table)
print(f"Chi-squared statistic: {chi2_stat}")
print(f"P-value: {p_value}")
alpha = 0.05
if p_value < alpha:
    print("Có đủ bằng chứng để bác bỏ H0.")
else:
    print("Không đủ bằng chứng để bác bỏ H0.")
```

Output:

```
Chi-squared statistic: 54.30602007353474
P-value: 1.6128633524576768e-12
Có đủ bằng chứng để bác bỏ H0.
```

Nhận xét: Có thể thấy được rằng trị giá P-value rất nhỏ. Điều này cho thấy được có sự chênh lệch trong tỷ trọng mắc bệnh về giấc ngủ giữa nam và nữ, và điều này hoàn toàn đúng với nhận định ban đầu của nhóm.

5.2.2 BMI ảnh hưởng tới bệnh về giấc ngủ:

Tiếp theo là chỉ số BMI, đây là một chỉ số sức khỏe phổ biến mà qua đó người ta đánh giá sơ bộ được một người có đang mất cân bằng giữa chiều cao và cân nặng hay không. Đây cũng là một chỉ số mà là hệ quả của nhiều bệnh tật khác trên cơ thể biểu hiện ra bên ngoài. Vậy hãy xem trong bộ dữ liệu này những người có chỉ số BMI khác nhau có sự khác nhau trong tình trạng bệnh về giấc ngủ hay không.

Giả thuyết: Chỉ số BMI và bệnh về giấc ngủ là không liên quan tới nhau.

H0: Chỉ số BMI và các bệnh về giấc ngủ là ĐỘC LẬP

H1: Chỉ số BMI và các bệnh về giấc ngủ là PHỤ THUỘC lẫn nhau

```
cross_tab = pd.crosstab(df['BMI Category'], df['Sleep Disorder'])
new_df = cross_tab.reset_index()
new_df.columns.name = None
new_df
```

Output:

	BMI Category	Insomnia	None	Sleep Apnea
0	Normal	9	200	7
1	Obese	4	0	6
2	Overweight	64	19	65

```
##-----------------------------------------------------
## Các giả thuyết kiểm định
##    H0: Chỉ số BMI và rối loạn giấc ngủ là ĐỘC LẬP
##    Ha: Chỉ số BMI và rối loạn giấc ngủ là PHỤ THUỘC lẫn nhau
##-----------------------------------------------------
alpha            = .05
confidence_level = (1-alpha)
contingency_table = pd.crosstab(df['BMI Category'], df['Sleep
Disorder'])
chi2_stat, p_value, _, _ = chi2_contingency(contingency_table)

print(f"Chi-squared statistic: {chi2_stat}")
print(f"P-value: {p_value}")

alpha = 0.05
if p_value < alpha:
    print("Có đủ bằng chứng để bác bỏ H0.")
else:
    print("Không đủ bằng chứng để bác bỏ H0.")
```

Output:

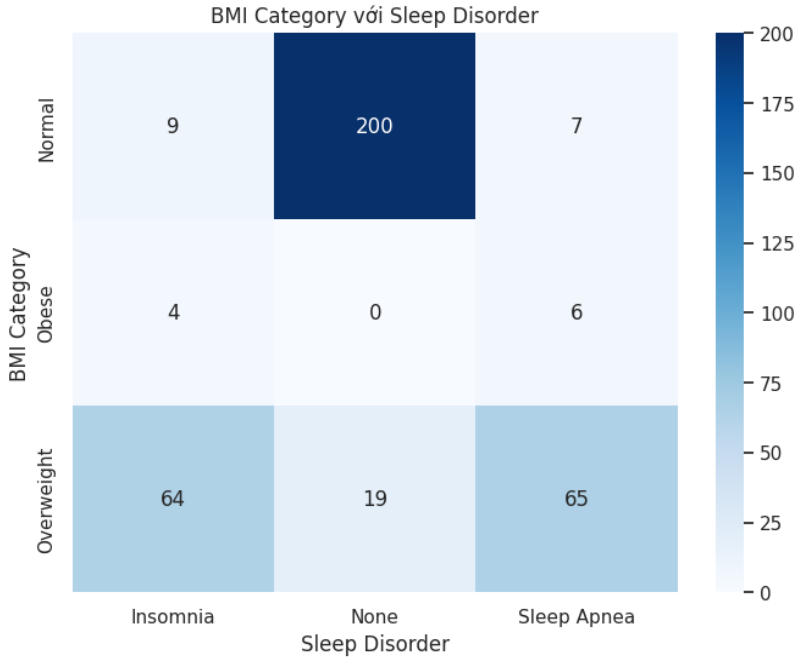
```
Chi-squared statistic: 245.66534355746683
P-value: 5.5883512097923584e-52
Có đủ bằng chứng để bác bỏ H0.
```

```
stat, p, dof, expected = stats.chi2_contingency(table)
## Kết luận theo phương pháp p-value (trị số p)
if (p < alpha):
    print(f'Trị số p = {p:.4f} < {alpha:.4f} nên bác bỏ H0 ==> (Chỉ số BMI, Rối loạn giấc ngủ) PHỤ THUỘC')
else:
    print(f'Trị số p = {p:.4f} >= {alpha:.4f} KHÔNG bác bỏ H0 ==> (Chỉ số BMI, Rối loạn giấc ngủ) ĐỘC LẬP')
```

Output:

```
Trị số p = 0.0000 < 0.0500 nên bác bỏ H0 ==> (Chỉ số BMI, Rối loạn giấc ngủ) PHỤ THUỘC
```

Ở đây kết quả trả về là bác bỏ H0, chứng tỏ rằng chỉ số BMI và các bệnh về giấc ngủ phụ thuộc lẫn nhau. Và hãy biểu diễn lại thông số này qua biểu đồ để có cái nhìn rõ ràng hơn về sự phụ thuộc này.



Nhận xét:

- Những người mắc chứng béo phì có xu hướng mắc chứng ngưng thở và chứng mất ngủ hơn người bình thường. Khi ở đây nhóm thấy rõ được rằng những người có trạng thái BMI là “Obese” hầu như 100% mắc bệnh (trong tập dữ liệu này). Và với người

thừa cân “Overweight” thì cũng có các trường hợp mắc bệnh về giấc ngủ cao hơn rất nhiều so với những người “Normal”.

- Những người có chỉ số BMI bình thường hiếm khi mắc các bệnh về giấc ngủ.

5.2.3 Độ tuổi và bệnh về sức khỏe:

Như lúc nãy tác giả đã nhận xét ở trên, từ khoảng độ tuổi 40 trở lên thì tỉ lệ mắc bệnh về giấc ngủ tăng đột biến hơn nhóm còn lại. Vậy hãy cùng đặt ra giả thuyết và đi kiểm định vấn đề này.

Giả thuyết: Những người trên 40 tuổi có khả năng mắc các bệnh về giấc ngủ cao hơn so với những người còn lại:

$$H0: \mu\{Age > 40\}[Sleep\ Disorder] = \mu\{Age =< 40\}[Sleep\ Disorder]$$

$$H1: \mu\{Age > 40\}[Sleep\ Disorder] \neq \mu\{Age =< 40\}[Sleep\ Disorder]$$

```
import pandas as pd
from statsmodels.stats.weightstats import ztest

# Tạo hai nhóm dữ liệu
group_over_40 = df[df['Age'] > 40]['Sleep_Disorders']
group_under_40 = df[df['Age'] <= 40]['Sleep_Disorders']

# Thực hiện kiểm định t-test
t_statistic, p_value = ztest(group_over_40, group_under_40, value=0,
alternative='two-sided')

# In kết quả
print(f'T-statistic: {t_statistic}\nP-value: {p_value}')

# Kiểm tra giả thuyết
alpha = 0.05
if p_value < alpha:
    print("Có bằng chứng để bác bỏ giả thiết null. Có sự khác biệt
đáng kể giữa hai nhóm.")
else:
    print("Không đủ bằng chứng để bác bỏ giả thiết null. Không có sự
khác biệt đáng kể giữa hai nhóm.")
```

Output:

```
T-statistic: 11.345336369021489
P-value: 7.8224860663594e-30
Có bằng chứng để bác bỏ giả thiết null. Có sự khác biệt đáng kể giữa
```

```
hai nhóm.
```

Nhận xét: Kết quả sau khi chạy kiểm định cho thấy có đủ bằng chứng để bác bỏ giả thuyết H0 và có sự khác biệt đáng kể giữa hai nhóm.

- T-statistic: Trong trường hợp này, giá trị 11.35 là khá đáng kể, cho thấy sự khác biệt giữa tỉ lệ mắc bệnh về giấc ngủ ở nhóm trên 40 tuổi và nhóm dưới hoặc bằng 40 tuổi là đáng kể.
- P-value: P-value rất thấp, điều này cho thấy được sự khác biệt giữa 2 nhóm là đáng kể.

Vì vậy, dựa vào kết quả này, tác giả có thể kết luận rằng tỉ lệ mắc bệnh về giấc ngủ ở nhóm trên 40 tuổi khác biệt đáng kể so với nhóm dưới hoặc bằng 40 tuổi. Cụ thể hơn, dựa vào điểm số T-statistic dương thì có thể cho rằng tỉ lệ mắc bệnh về giấc ngủ ở nhóm trên 40 tuổi có vẻ cao hơn so với nhóm dưới hoặc bằng 40 tuổi.

5.2.4 Nhịp tim và bệnh về giấc ngủ:

Như ở phần trên nhóm đã đưa ra nhận định được rằng những người có nhịp tim cao (từ 78 trở lên) sẽ có tỉ lệ bị bệnh về giấc ngủ cao hơn những người có nhịp tim thấp hơn. Và lần này hãy thử kiểm định lại trên toàn bộ tập các giá trị nhịp tim để xem rằng nhịp tim và các bệnh về giấc ngủ có liên quan đến nhau hay không.
Giả thuyết: Nhịp tim và tỉ lệ mắc các bệnh về giấc ngủ là không liên quan đến nhau.
Cách thức: Chia thành 2 nhóm dữ liệu, một là những nhóm nhịp tim không bị bệnh về giấc ngủ và nhóm còn lại là bị bệnh về giấc ngủ.

```
df['Sleep_Disorderss'] = df['Sleep Disorder'].replace({
    'None': 'Không',
    'Insomnia': 'Có',
    'Sleep Apnea': 'Có',})
```

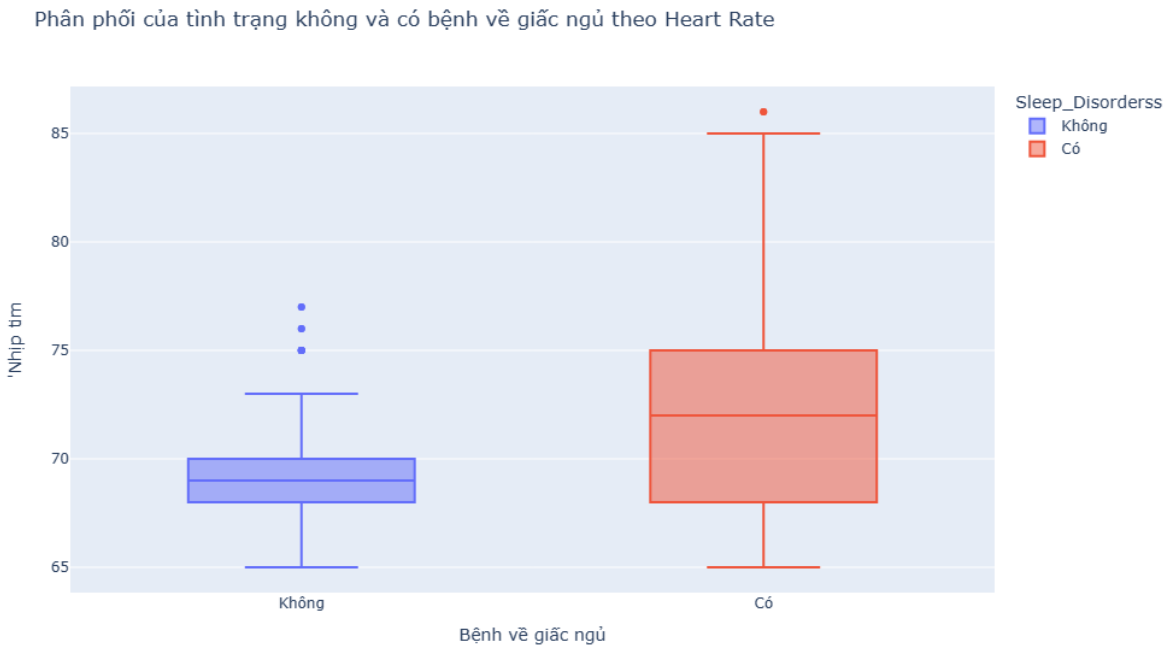
```
# Tạo hai nhóm dữ liệu
group_none = df[df['Sleep_Disorderss'] == 'Không']['Heart Rate']
group_not_none = df[df['Sleep_Disorderss'] == 'Có']['Heart Rate']
# Thực hiện kiểm định t-test
t_statistic, p_value = ztest(group_none, group_not_none, value=0,
alternative='two-sided')
# In kết quả
print(f'T-statistic: {t_statistic}\nP-value: {p_value}')

# Kiểm tra giả thuyết
alpha = 0.05
```

```
if p_value < alpha:
    print("Có bằng chứng để bác bỏ giả thiết null. Có sự khác biệt đáng kể giữa hai nhóm.")
else:
    print("Không đủ bằng chứng để bác bỏ giả thiết null. Không có sự khác biệt đáng kể giữa hai nhóm.")
```

Output:

```
T-statistic: -5.60803984488608
P-value: 2.0463092339913586e-08
Có bằng chứng để bác bỏ giả thiết null. Có sự khác biệt đáng kể giữa hai nhóm.
```

Qua kết quả chạy kiểm định cho thấy được có sự khác biệt đáng kể giữa hai nhóm. Điều này có nghĩa rằng tỉ lệ mắc bệnh về giấc ngủ theo nhịp tim là có sự khác biệt đáng kể. Để rõ hơn nhóm sẽ vẽ biểu đồ để trực quan.



Nhận xét:

- Qua biểu đồ và kết quả kiểm định cho thấy được những người có nhịp tim cao từ khoảng trên 70 có khả năng bị bệnh về giấc ngủ cao hơn tập còn lại. Và với những người từ 80 nhịp tim đổ lên thì trong bộ dữ liệu này ai cũng bị bệnh về nhịp tim cả.

5.2.5 Thời gian thể dục trong ngày với bệnh về giấc ngủ:

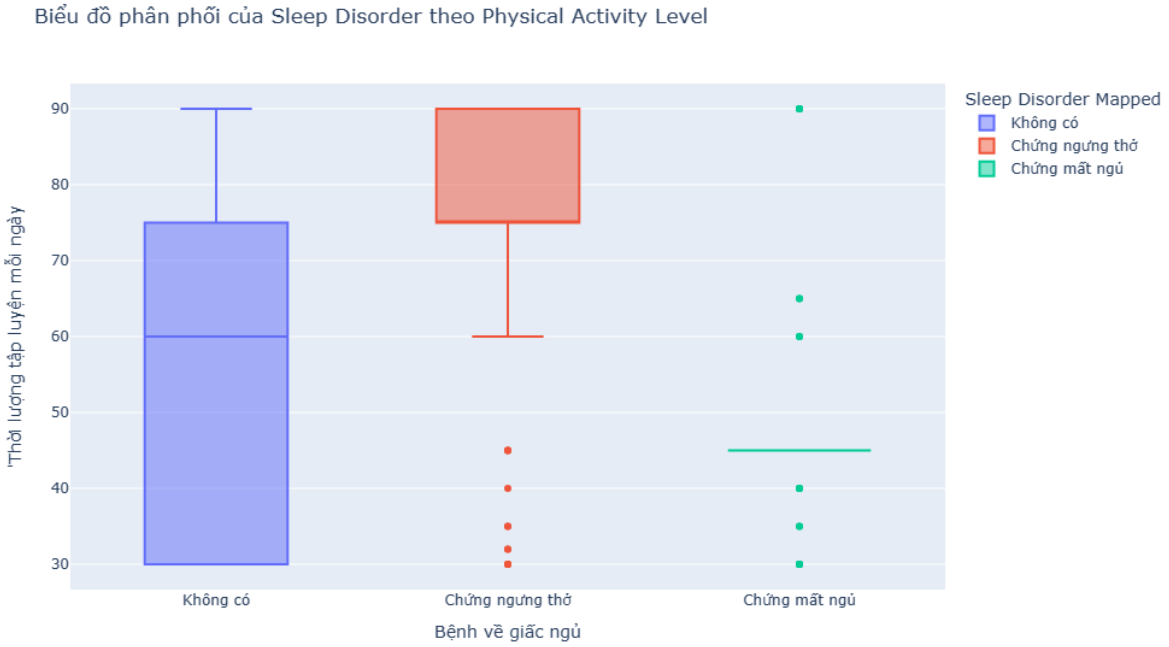
Sau khi nhóm đã đánh giá các yếu tố sinh học ảnh hưởng đến các vấn đề về giấc ngủ, bây giờ nhóm sẽ điều tra xem lối sống tập thể dục của những người tham gia khảo sát có ảnh hưởng như thế nào đối với tình trạng sức khỏe của họ, cụ thể hơn là trong việc mắc các bệnh về giấc ngủ.

```
# Tạo mô hình OLS
model = sm.OLS.from_formula('Q("Physical Activity Level") ~
C(Sleep_Disorder)', data=df).fit()
residuals = model.resid
# Thực hiện kiểm định Levene trên residuals
levene_statistic, levene_p_value =
levene(*[residuals[df['Sleep_Disorder'] == occupation]
                                        for occupation in
df['Sleep_Disorder'].unique()])
print("Levene Statistic:", levene_statistic)
print("P-value (Levene):", levene_p_value)

alpha = 0.05
if levene_p_value < alpha:
    print("Có bằng chứng để bác bỏ giả định về phương sai đồng
nhất.")
else:
    print("Không có bằng chứng để bác bỏ giả định về phương sai đồng
nhất.")
```

Output:

```
Levene Statistic: 26.637945259436496
P-value (Levene): 1.549139850268813e-11
Có bằng chứng để bác bỏ giả định về phương sai đồng nhất.
```

Ở đây kiểm định Levene cho thấy có bằng chứng để bác bỏ giả định về phương sai đồng nhất (sự đồng nhất về phương sai giữa các nhóm không được đảm bảo), vậy nhóm không dùng kiểm định Anova ở đây vì trong trường hợp này kết quả kiểm định không đáng tin cậy.
Tuy nhiên vẫn còn cách khác để tác giả nhận định vấn đề này đó là việc trực quan hóa lên:



Nhận xét: Qua biểu đồ cho thấy được những những người mang hội chứng "Sleep Apnea" thường xuất hiện ở những cá nhân được khảo sát có mức độ tập luyện mỗi ngày cao (từ 80-90 phút). Trong khi đó những người tập thể dục đều đặn và vừa phải chủ yếu từ 30-70 phút mỗi ngày thì hầu như không gặp vấn đề về giấc ngủ, ngoại trừ một số ít mắc 2 bệnh còn lại nhưng so với tổng thể thì không nhiều.

5.3 Các yếu tố còn lại:

5.3.1 Chỉ số BMI và tuổi tác:

Giả thuyết: Không có sự khác biệt giữa những nhóm người có chỉ số BMI khác nhau về tuổi tác.

```
model = sm.OLS.from_formula('Q("Age") ~ C(Q("BMI Category"))',
data=df).fit()
# Lấy residuals
residuals = model.resid
# Thực hiện kiểm định Levene trên residuals
levene_statistic, levene_p_value = levene(*[residuals[df['BMI
Category'] == category] for category in df['BMI Category'].unique()])

print("Levene Statistic:", levene_statistic)
print("P-value (Levene):", levene_p_value)
alpha = 0.05
if levene_p_value < alpha:
```

```
    print("Có bằng chứng để bác bỏ giả định về phương sai đồng
nhất.")
else:
    print("Không có bằng chứng để bác bỏ giả định về phương sai đồng
nhất.")
```

Output:

```
Levene Statistic: 1.5485012024226694
P-value (Levene): 0.21393699486209478
Không có bằng chứng để bác bỏ giả định về phương sai đồng nhất.
```

```
from scipy.stats import f_oneway
# Ví dụ với Age và Occupation
grouped_data = [df['Age'][df['BMI Category'] == occupation] for
occupation in df['BMI Category'].unique()]
# Thực hiện kiểm định ANOVA
f_statistic, p_value = f_oneway(*grouped_data)
print("F-statistic:", f_statistic)
print("P-value:", p_value)
alpha = 0.05
if p_value < alpha:
    print("Có đủ bằng chứng để bác bỏ H0.")
else:
    print("Không đủ bằng chứng để bác bỏ H0.")
```

Output:

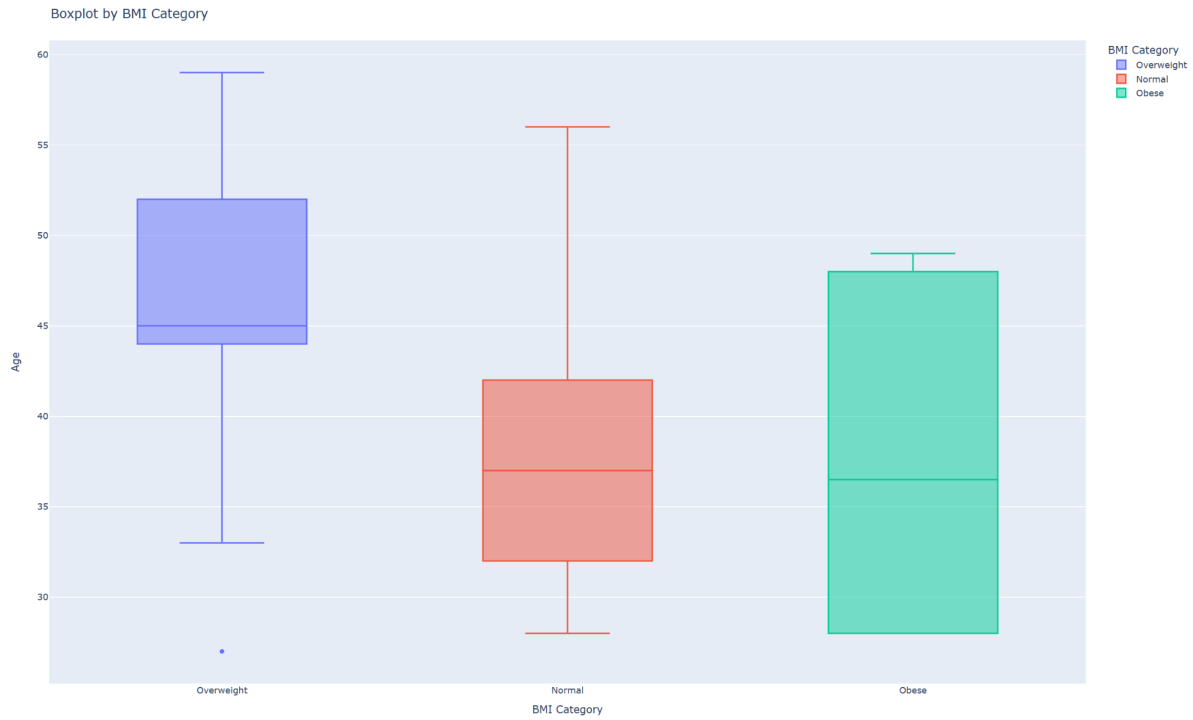
```
F-statistic: 73.48591878915192
P-value: 1.3033764297692233e-27
Có đủ bằng chứng để bác bỏ H0.
```

- Hậu kiểm Tukey HSD.

```
from statsmodels.stats.multicomp import pairwise_tukeyhsd
data_for_anova = pd.DataFrame({'Age': df['Age'], 'BMI_Category':
df['BMI Category']})
# Thực hiện hậu kiểm Turkey
tukey_result = pairwise_tukeyhsd(endog=data_for_anova['Age'],
groups=data_for_anova['BMI_Category'], alpha=0.05)
# In kết quả hậu kiểm Turkey
print(tukey_result)
```

Output:

```
 Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
group1    group2    meandiff p-adj   lower  upper  reject
-----------------------------------------------------
Normal      Obese  -0.4722 0.9785 -6.0744 5.1299  False
Normal Overweight   9.4129    0.0  7.5648 11.261   True
 Obese Overweight   9.8851 0.0001  4.2263 15.544   True
-----------------------------------------------------
```



Nhận xét: Qua biểu đồ và các kết quả kiểm định cho thấy được rằng những người có chỉ số BMI là Normal và Obese có sự phân phối độ tuổi là như nhau. Trong khi những người có độ tuổi cao thì khả năng bị thừa cân "Overweight" tăng lên đáng kể.

5.3.2 Số bước chân đi trong ngày và số lần thức giấc ban đêm:

Giả thuyết: Số bước đi trong ngày càng tăng thì số lần thức giấc ban đêm càng giảm.

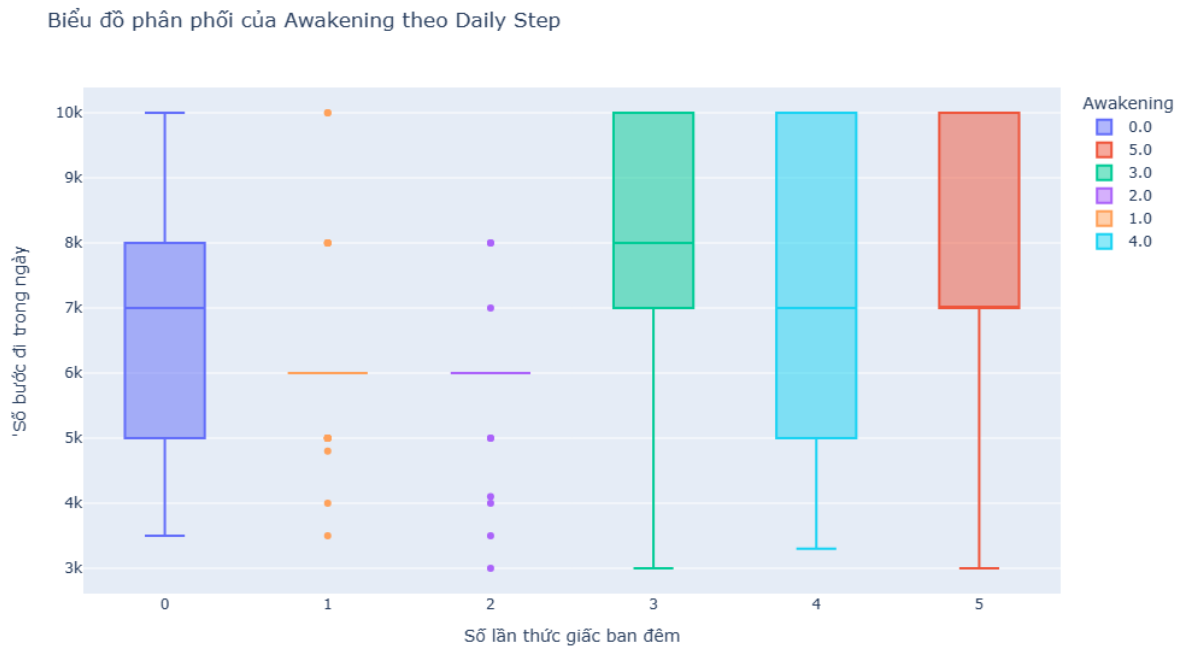
- Tính hệ số tương quan Pearson.

```
## Tính hệ số tương quan Pearson
r, _ = stats.pearsonr(df['Daily Steps'], df['Awakening'])
r = round(r, 2)
if np.sign(r) == -1:
    s = "âm"
elif np.sign(r) == 1:
    s = "dương"
if np.abs(r) < 0.1:
  print('Số bước đi trong ngày và số lần tỉnh giấc trong 1 đêm không có mối quan hệ tương quan tuyến tính')
elif np.abs(r) >= 0.1 and np.abs(r) < 0.3:
  print('Số bước đi trong ngày và số lần tỉnh giấc trong 1 đêm có mối tương quan tuyến tính',s, 'rất yếu')
elif np.abs(r) >= 0.3 and np.abs(r) < 0.5:
  print('Số bước đi trong ngày và số lần tỉnh giấc trong 1 đêm có mối tương quan tuyến tính',s, 'trung bình')
else:
  print('Số bước đi trong ngày và số lần tỉnh giấc trong 1 đêm có mối tương quan tuyến tính',s, 'mạnh')
```

Output:

```
Số bước đi trong ngày và số lần tỉnh giấc trong 1 đêm có mối tương quan tuyến tính dương rất yếu
```

Ở đây, hệ số Pearson cho thấy số bước đi trong ngày và số lần tỉnh giấc ban đêm có mối quan hệ tuyến tính dương rất yếu. Để rõ hơn nhóm sẽ vẽ biểu đồ để trực quan.



Nhận xét: Qua biểu đồ, ta có thể thấy rằng những người đi bộ nhiều mỗi ngày có xu hướng dễ tỉnh giấc vào ban đêm hơn so với những người ít đi bộ. Điều này trái với giả thuyết ban đầu nhóm đưa ra. Nguyên nhân có thể lý giải cho kết luận trên là mặc dù việc đi bộ có rất nhiều lợi ích, nhưng đi bộ quá nhiều có thể khiến cơ thể mệt mỏi cũng như gây mất nước, khiến mọi người phải thức dậy vào ban đêm để uống nước.

5.3.3 Mức độ hoạt động thể chất và số lần thức giấc ban đêm

Giả thuyết: Thời gian hoạt động thể chất càng tăng thì số lần thức giấc ban đêm sẽ tăng theo.

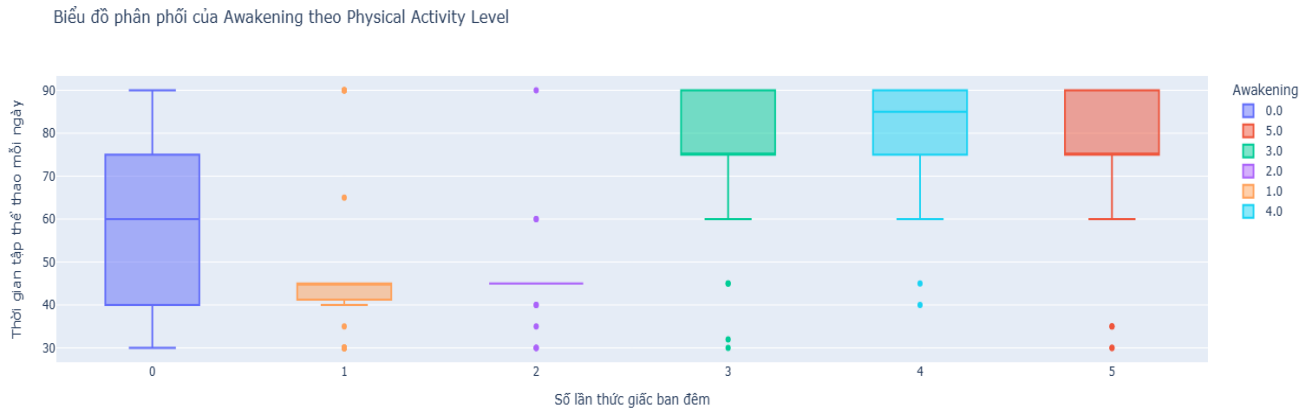
- Tính hệ số tương quan Pearson.

```
## Tính hệ số tương quan Pearson
r, _ = stats.pearsonr(df['Physical Activity Level'], df['Awakening'])
r = round(r, 2)
if np.sign(r) == -1:
    s = "âm"
elif np.sign(r) == 1:
    s = "dương"
if np.abs(r) < 0.1:
  print('Mức độ hoạt động thể chất và số lần tỉnh giấc trong 1 đêm
không có mối quan hệ tương quan tuyến tính')
elif np.abs(r) >= 0.1 and np.abs(r) < 0.3:
  print('Mức độ hoạt động thể chất và số lần tỉnh giấc trong 1 đêm có
mối tương quan tuyến tính',s, 'rất yếu')
elif np.abs(r) >= 0.3 and np.abs(r) < 0.5:
  print('Mức độ hoạt động thể chất và số lần tỉnh giấc trong 1 đêm có
mối tương quan tuyến tính',s, 'trung bình')
else:
  print('Mức độ hoạt động thể chất và số lần tỉnh giấc trong 1 đêm có
mối tương quan tuyến tính',s, 'mạnh')
```

Output:

```
Mức độ hoạt động thể chất và số lần tỉnh giấc trong 1 đêm có mối
tương quan tuyến tính dương rất yếu
```

Vì 2 biến trên có mối quan hệ tương quan tuyến tính rất yếu, nghĩa là giả thuyết ban đầu nhóm đưa ra là đúng. Nhóm sẽ trực quan hóa dữ liệu để dễ dàng đưa ra nhận xét.



Nhận xét: Việc tập thể dục ở cường độ cao có thể dẫn tới việc bị thức giấc vào ban đêm diễn ra thường xuyên hơn. Trong khi đó, những người tập thể dục với thời lượng tập vừa phải hầu như không bị thức giấc vào ban đêm. Vì vậy, để có được giấc ngủ sâu, người tập luyện nên xây dựng lịch trình tập luyện phù hợp với sức khỏe của cơ thể.

- **Tổng kết:**

Dựa trên kết quả kiểm định các yếu tố ảnh hưởng đến thời lượng và chất lượng giấc ngủ, tác giả nhận thấy rằng giới tính, huyết áp, BMI, độ tuổi, nhịp tim, và thời gian tập thể dục đều đóng vai trò quan trọng trong tình trạng giấc ngủ. Các liên kết phức tạp giữa các yếu tố này và bệnh về giấc ngủ đã được chỉ ra, đặt ra những thách thức và cơ hội trong việc quản lý sức khỏe giấc ngủ.

Qua những nhận định này, tác giả hiểu rằng việc xây dựng một chiến lược chăm sóc sức khỏe giấc ngủ toàn diện không chỉ đơn thuần là điều chỉnh yếu tố giới tính, mà còn đòi hỏi sự đa chiều hóa, tính đồng bộ với huyết áp, BMI, độ tuổi, nhịp tim, và thói quen vận động. Điều này là cực kỳ quan trọng để cải thiện chất lượng cuộc sống và ngăn chặn bệnh về giấc ngủ.

Tuy nhiên, những nhận định này chỉ là sự bắt đầu, và quá trình kiểm định và phân tích sẽ tiếp tục để đảm bảo rằng nhóm có một hiểu biết chi tiết và chính xác về ảnh hưởng của các yếu tố này. Điều này sẽ giúp xây dựng một chiến lược chăm sóc sức khỏe giấc ngủ đặc sắc, hiệu quả, và linh hoạt để đáp ứng đa dạng của cộng đồng và cá nhân.

CHƯƠNG VI. KHAI THÁC DỮ LIỆU NGHIÊN CỨU

6.1 Xây dựng mô hình dự báo:

6.2.1 Phân lớp dữ liệu:

Phân lớp dữ liệu là một kỹ thuật quan trọng trong lĩnh vực máy học, nó giúp tự động hóa quá trình gán nhãn cho dữ liệu dựa trên một tập huấn luyện trước. Quy trình phân lớp thường được thực hiện qua hai bước chính:

- **Bước 1:** Học (Training):

Trong bước này, mô hình phân lớp được xây dựng dựa trên tập dữ liệu huấn luyện. Mô hình sẽ học các quy tắc và mối quan hệ trong dữ liệu để có thể phân lớp đối tượng vào các nhóm hoặc lớp đã định nghĩa trước. Các thuật toán khác nhau như cây quyết định, máy vector hỗ trợ, hoặc mạng nơ-ron có thể được sử dụng để xây dựng mô hình phân lớp.

- **Bước 2:** Kiểm tra và Đánh giá (Testing and Evaluation):

Sau khi mô hình đã được xây dựng, nó được đặt vào thử nghiệm bằng cách sử dụng dữ liệu mới hoặc dữ liệu chưa được sử dụng trong quá trình huấn luyện. Mô hình sẽ phân loại các đối tượng vào các lớp tương ứng. Quá trình này thường được đánh giá thông qua các độ đo hiệu suất như độ chính xác, độ nhạy, độ đặc biệt, và ma trận lỗi.

6.2.2 Mục đích phân lớp cho bộ dữ liệu

Với sự nhận thức về tầm quan trọng của giới tính, huyết áp, BMI, độ tuổi, nhịp tim, và các yếu tố lối sống khác ảnh hưởng tới sức khỏe giấc ngủ, nhóm quyết định tiếp tục nghiên cứu bằng cách xây dựng một mô hình dự báo. Mô hình này sẽ được thiết kế để dự đoán khả năng xuất hiện bệnh về giấc ngủ dựa trên những thông tin đã được thu thập.

Với việc biến dự báo là “Sleep Disorder” chỉ mang 3 giá trị là “None”, “Insomnia” và”Sleep Apnea” thì mô hình dự báo nhóm sẽ thể sử dụng các thuật toán quen thuộc và phù hợp đã được giảng dạy trong chương trình đào tạo như K-Nearest Neighbors, Logistic Regression, Random Forest, và Decision Tree. Những thuật toán khác như Linear Regression hay Naive Bayes có thể không phù hợp trong trường hợp này. Và sau đó đánh giá xem những mô hình nào cho kết quả tốt nhất để phân tích và đánh giá.

Và từ đó, nhóm hy vọng rằng mô hình dự báo sẽ không chỉ cung cấp thông tin quan trọng về nguy cơ bệnh về giấc ngủ, mà còn giúp tạo ra các chiến lược chăm sóc cá nhân hóa và hiệu quả. Quá trình này sẽ mang lại cơ hội để ứng dụng những phát hiện trong nghiên cứu vào thực tế và định hình các chương trình quản lý sức khỏe giấc ngủ tương lai.

6.2.3 Xây dựng mô hình phân lớp:

a. Tiền xử lý dữ liệu trước khi phân lớp:

Với những thuật toán phân lớp và hồi quy như K-Nearest Neighbors, Logistic Regression, Random Forest, và Decision Tree thường cần dữ liệu đầu vào ở dạng số. Do đó, tác giả sẽ kiểm tra các biến categorical trong dữ liệu, và xem xét mã hóa nó bằng các phương pháp như One-hot Encoding, Label Encoder,... để dữ liệu có thể phù hợp hơn khi đưa vào huấn luyện mô hình. Dưới đây là những cột có Categorical của bộ dữ liệu:

```
for col in df.select_dtypes(include=['object']).columns:
    print(f"{col}: {df[col].unique()}")
```

Output:

```
Gender: ['Male' 'Female']
Occupation: ['Software Engineer' 'Doctor' 'Sales Representative'
'Teacher' 'Nurse'
 'Engineer' 'Accountant' 'Scientist' 'Lawyer' 'Salesperson'
'Manager']
BMI Category: ['Overweight' 'Normal' 'Obese']
Sleep Disorder: ['None' 'Sleep Apnea' 'Insomnia']
```

Nhận thấy rằng: Có thể chia thành 2 nhóm để mã hóa:

- Gender, Sleep Disorder:

Với biến giới tính có chỉ hai giá trị ('Male' và 'Female'), hoàn toàn có thể sử dụng Label Encoder vì nó chỉ có hai giá trị và không tạo ra mối quan hệ tuần tự. Với biến bệnh về giấc ngủ "Sleep Disorder", cũng có thể sử dụng Label Encoder nếu có thứ bậc giữa các loại rối giấc ngủ.

- Occupation, BMI Category:

Với biến nghề nghiệp, có nhiều giá trị khác nhau. Mã hóa One-hot có thể là lựa chọn tốt để không tạo ra thứ bậc giữa các nghề nghiệp và giữ cho mô hình không giả định về mối quan hệ tuần tự. Tương tự, nhóm không muốn tạo mối quan hệ tuần tự giữa các chỉ số BMI (như 'Normal' < 'Overweight' < 'Obese'), vậy nên mã hóa One-hot là một lựa chọn hợp lý.

Tiến hành mã hóa:

```
# Mã hóa cột "Occupation" bằng one-hot encoding
df = pd.get_dummies(df, columns=['Occupation'], prefix='Occupation')
df = pd.get_dummies(df, columns=['BMI Category'], prefix='BMI
Category')
print(df.head())
```

```
# Mã hóa những cột còn lại bằng Label Encoder
label_encoder = LabelEncoder()
df['Gender'] = label_encoder.fit_transform(df['Gender'])
df['Sleep Disorder'] = label_encoder.fit_transform(df['Sleep
Disorder'])
```

b. Chia tập Train và Test của bộ dữ liệu:

Ở đây vì bộ dữ liệu có lượng mẫu khá nhỏ nên vì thế để đảm bảo việc tập test có đủ không gian để đánh giá sau khi chạy các mô hình, nhóm quyết định sẽ chia theo tỉ lệ 70-30, 70% dùng để huấn luyện và 30% dùng để chạy kết quả.

```
X = df.drop(['Sleep Disorder'], axis=1)
y = df['Sleep Disorder']

# Chia thành các tập Train và Test
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3, random_state=10)
```

Sử dụng Pipeline giúp chạy các khâu chuẩn hóa dữ liệu về các định dạng hợp lý. Và sau đó khởi tạo các tham số mẫu để có sau này dùng thư viện Grid_search tìm ra những tham số phù hợp cho mô hình.

```
pipeline = Pipeline([
    ('scaler', StandardScaler()),
    ('clf', RandomForestClassifier())])
```

```
param_grid = [
    {'clf': [RandomForestClassifier()],
        'clf__n_estimators': [100, 200, 300,400],
        'clf__max_depth': [None, 5, 10,15],},

    {'clf': [LogisticRegression()],
        'clf__solver': ['liblinear', 'lbfgs'],
        'clf__C': [0.01,0.1, 1, 10],},

    {'clf': [KNeighborsClassifier()],
        'clf__n_neighbors': [3, 5, 7,9],},

    {'clf': [DecisionTreeClassifier()],
        'clf__max_depth': [None, 5, 10,15],}]
```

Giải thích:

`param_grid` mà bạn đã cung cấp là một lưới tham số được sử dụng trong quá trình tinh chỉnh siêu tham số bằng cách sử dụng `GridSearchCV` từ thư viện scikit-learn. Lưới này

được xây dựng để tinh chỉnh bốn mô hình phân loại khác nhau, mỗi mô hình có các tham số tương ứng cần được điều chỉnh. Dưới đây là giải thích chi tiết:

- RandomForestClassifier:
  - 'clf__n_estimators': [100, 200, 300, 400] - Các mẫu thử số cây quyết định trong rừng (n_estimators).
  - 'clf__max_depth': [None, 5, 10, 15] - Các thông số về độ sâu tối đa của cây quyết định (max_depth) sẽ tính toán.

- LogisticRegression:
  - 'clf__solver': ['liblinear', 'lbfgs'] - Thuật toán sẽ được tính toán để chọn một trong 2 thuật toán giải quyết phù hợp cho mô hình (liblinear hoặc lbfgs).
  - 'clf__C': [0.01, 0.1, 1, 10]- Thử nghiệm để tìm ra giá trị hiệu chỉnh đồng chuẩn (C) tối ưu cho mô hình Logistic Regression

- KNeighborsClassifier:
  - 'clf__n_neighbors': [3, 5, 7, 9] - Tìm kiếm số lượng láng giềng được sử dụng để phù hợp với mô hình.

- DecisionTreeClassifier:
  - 'clf__max_depth': [None, 5, 10, 15] - Tìm kiếm độ sâu tối đa phù hợp của cây quyết định (max_depth).

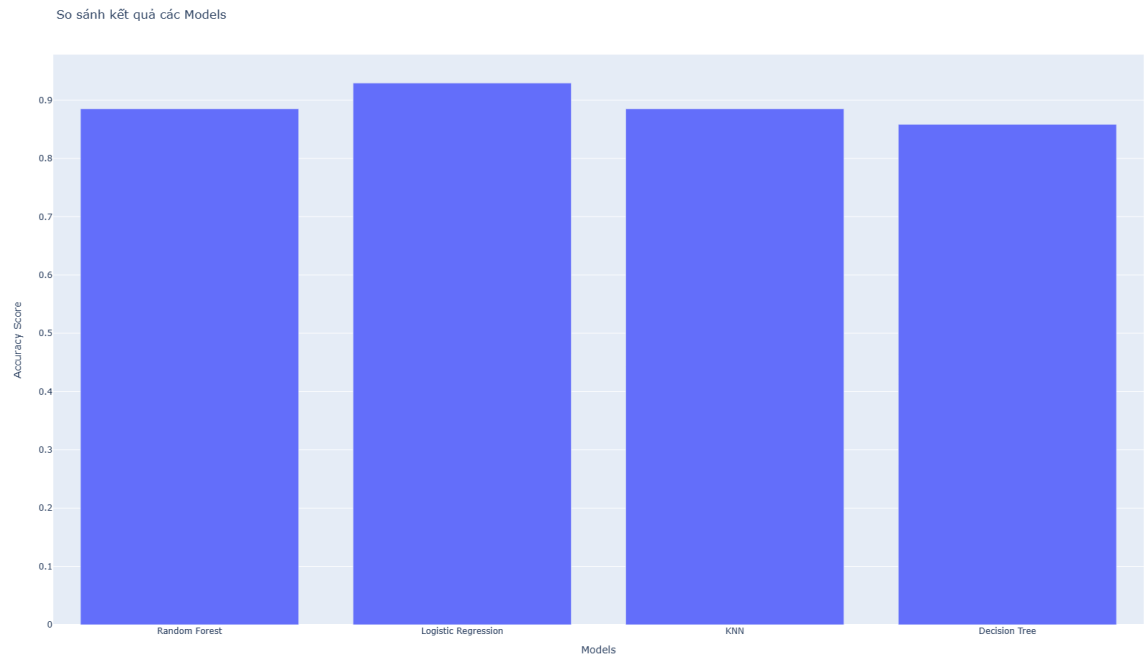
Những thay đổi trong các tham số này sẽ được thử nghiệm để tìm ra giá trị tối ưu giúp mô hình hoạt động hiệu quả nhất trên dữ liệu huấn luyện.

Và sau khi đã thiết lập xong các dữ kiện cần thiết, nhóm sẽ chạy 4 mô hình đã nói từ trước và đưa ra thông cố Accuracy để xem hiệu suất mỗi mô hình.

```
# Sử dụng GridSearchCV để bổ sung các tham số
grid_search = GridSearchCV(pipeline, param_grid, cv=5)
grid_search.fit(X_train, y_train)

# Lưu lại Best_model sau khi chạy GridSearch
best_model = grid_search.best_estimator_

# Chạy các thuật toán để tính Accuracy
models = [
    ('Random Forest', RandomForestClassifier()),
    ('Logistic Regression', LogisticRegression()),
    ('KNN', KNeighborsClassifier()),
    ('Decision Tree', DecisionTreeClassifier())]
```

```
accuracy_scores = []
for name, model in models:
    pipeline = Pipeline([
        ('scaler', StandardScaler()),
        ('clf', model)])
    pipeline.fit(X_train, y_train)
    y_pred = pipeline.predict(X_test)
    accuracy = accuracy_score(y_test, y_pred)
    accuracy_scores.append(accuracy)
```

Kết quả:



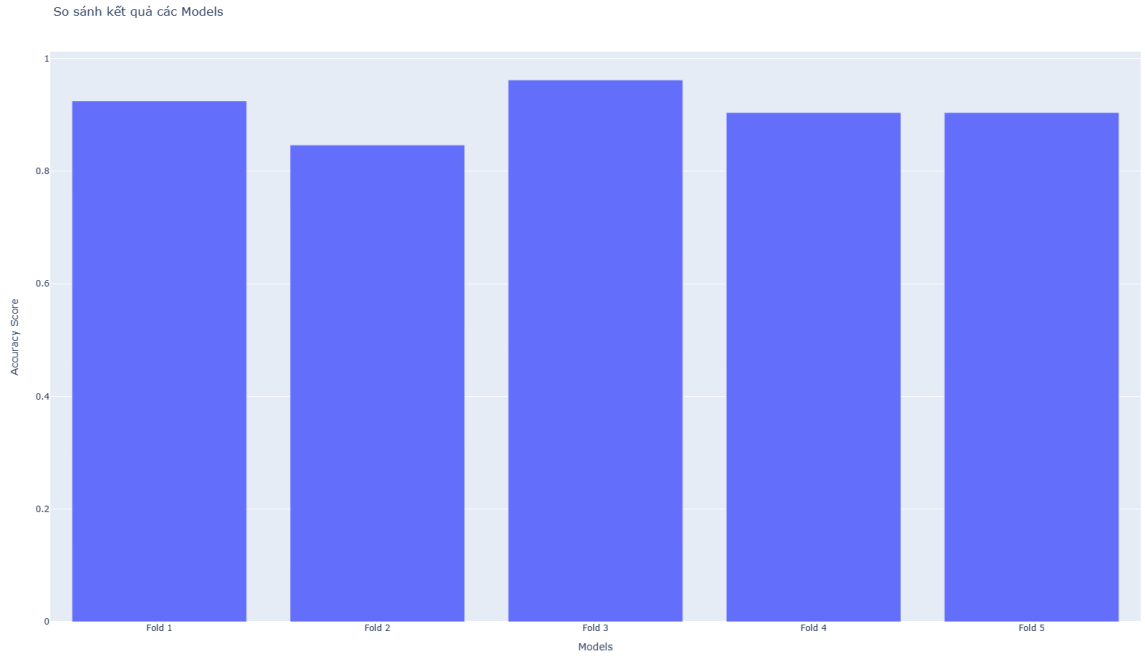
Tổng quan: Mô hình Logistic Regression khi có điểm số Accuracy trên 90% và là kết quả cao nhất, và theo sau đó là các mô hình còn lại với mức đánh giá chuẩn xác cũng khá cao tuy nhiên chưa tới mức 90%. Tuy nhiên vì đây là mô hình dự báo người có thể bị những bệnh về giấc ngủ hay không thì mỗi chỉ số Accuracy là chưa đủ, vậy nên hãy đến phần đánh giá để xem Logistic Regression có thật sự hiệu quả hay không.

c. Đánh giá mô hình máy học:

Với kết quả từ phần trên, nhóm sẽ quyết định sử dụng mô hình có số điểm cao nhất là Logistic Regression để tiến hành các đánh giá xem liệu mô hình này có phù hợp cho việc dự báo bệnh về giấc ngủ dựa trên tập dữ liệu.

Đầu tiên, khi nhận thấy mô hình có ít mẫu nhưng hiệu suất các mô hình là gần và hơn 90%. Điều này khiến nhóm quyết định kiểm tra xem liệu có tình trạng Overfitting ở đây hay không.

```
from sklearn.model_selection import cross_val_score

cv_scores = cross_val_score(best_model.named_steps['clf'], X_train,
y_train, cv=5)
print("Cross-Validation Scores for Best Model:", cv_scores)
print("Mean CV Score for Best Model:", np.mean(cv_scores))
```

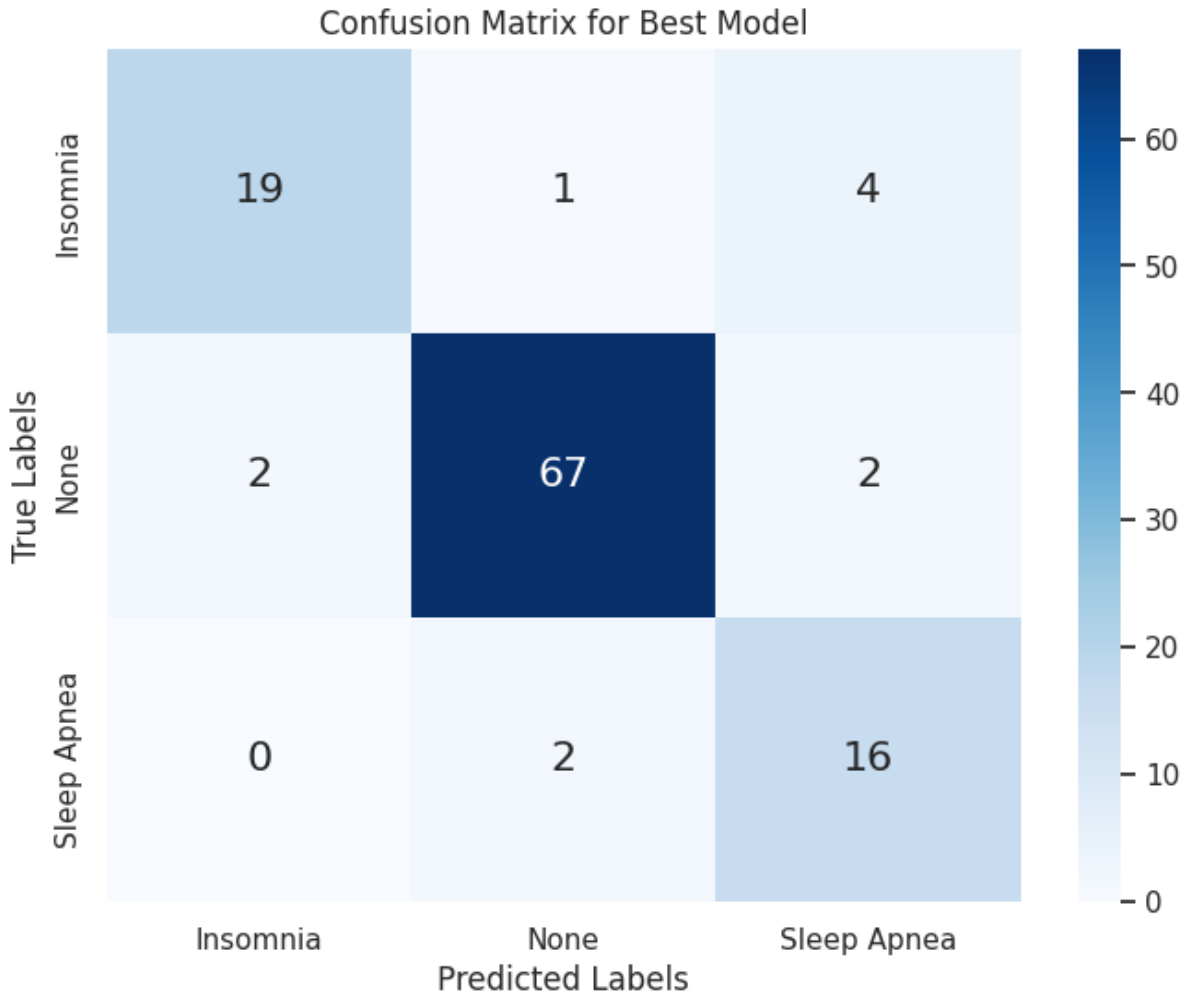



```
from sklearn.metrics import accuracy_score

y_pred_test = best_model.predict(X_test)
test_accuracy = accuracy_score(y_test, y_pred_test)
print("Accuracy on Test Set:", test_accuracy)

Output:

=> Accuracy on Test Set: 0.9026548672566371
```

Thông qua Cross-validation cho thấy được hiệu suất giữa những lần chạy là có sự thay đổi, khi các điểm số thay đổi từ khoảng 0.84 đến 0.96, nhưng giữa các lần chạy không có sự

chênh lệch lớn. Giá trị trung bình của các điểm số (Mean CV Score) là 0.9004, và sau đó thử trên tập test cho ra kết của 0.89 cho thấy mô hình có hiệu suất ổn định và không có dấu hiệu rõ ràng của overfitting.

Tiếp theo sẽ là sử dụng Ma trận nhầm lẫn để đánh giá được hiệu suất thực sự của mô hình:



	precision	recall	f1-score	support
0	0.90	0.79	0.84	24
1	0.96	0.94	0.95	71
2	0.73	0.89	0.80	18
accuracy			0.90	113
macro avg	0.86	0.87	0.86	113
weighted avg	0.91	0.90	0.90	113

Nhận xét:

Dựa trên kết quả đánh giá của mô hình phân loại, có thể thấy rằng mô hình có hiệu suất tốt tổng thể. Dưới đây là một số nhận xét về các chỉ số đánh giá:

- Insomnia (Nhãn 0):
    - Precision: 90% (Tỷ lệ số lượng dự đoán đúng Insomnia trên tổng số dự đoán là Insomnia)
    - Recall: 79% (Tỷ lệ số lượng dự đoán đúng Insomnia trên tổng số thực tế là Insomnia)
    - F1-score: 84% (Trung bình điều hòa giữa Precision và Recall)

- Sleep Apnea (Nhãn 2):
    - Precision: 73% (Tỷ lệ số lượng dự đoán đúng Insomnia trên tổng số dự đoán là Insomnia)
    - Recall: 89% (Tỷ lệ số lượng dự đoán đúng Insomnia trên tổng số thực tế là Insomnia)
    - F1-score: 80% (Trung bình điều hòa giữa Precision và Recall)

- Tổng thể:
    - Accuracy: 90% (Tỷ lệ số lượng dự đoán đúng trên tổng số mẫu)
    - Macro Avg (trung bình không trọng số của precision, recall và F1-score): 86%
    - Weighted Avg (trung bình có trọng số theo số lượng mẫu của precision, recall và F1-score): 90%

Về tổng quan, mô hình đưa ra kết quả khá tốt khi hiệu suất các mục Precision, Recall hay F1-Score đều từ 70% đổ lên. Và với mục đích của mô hình là dự báo được những người bị bệnh thì chỉ số Recall thực sự là quan trọng hơn những chỉ số khác. Khi thứ nhóm nghiên cứu cần là tối đa hóa được tỷ lệ dự đoán đúng bệnh và bỏ sót ít người bị bệnh nhất có thể. Vì khi dự đoán sai sẽ dẫn tới những sai lầm trong điều trị và có thể dẫn tới tình trạng tồi tệ hơn. Ở đây có thể thấy được hiệu suất dự báo những người bị "Sleep Apnea" là tốt khi đạt tới 89%. Trong khi đó ở Insomnia thì chỉ gần 80%.

6.3 Dự báo từ kết quả phân lớp dữ liệu đạt được:

Trong nghiên cứu này, nhóm chúng em đã phát triển một mô hình để dự đoán bệnh về giấc ngủ dựa trên tập dữ liệu được cung cấp. Kết quả cho thấy mô hình tối ưu nhất cho bộ dữ liệu này là Logistic Regression, khi có thể xác định được khá tốt những người bị "Insomnia" và đạt tương đối với những mẫu khảo sát có tình trạng "Sleep Apnea". Có thể do sự chênh lệch về số mẫu giữa các tình trạng bệnh về giấc ngủ nên hiệu suất dự báo còn chưa đạt kết quả tốt như mong đợi. Các biện pháp để có thể cải thiện mô hình như thu thập thêm lượng mẫu để thuật toán có thể học một cách sâu và nhiều hơn, để từ đó đưa ra kết quả dự báo tốt hơn hiện nay.
Thông qua dự báo trên, ta có thể thấy được dự báo này có rất nhiều công dụng đối với nhiều ngành nghề khác nhau như:

- Chẩn đoán và Điều trị: Xác định bệnh về giấc ngủ giúp đưa ra quyết định về liệu pháp và điều trị thích hợp.
- Dự báo Tương Lai: Dự báo tình trạng bệnh sẽ giúp dự đoán các vấn đề sức khỏe có thể phát sinh trong tương lai và thực hiện các biện pháp ngăn chặn kịp thời.
- Tư vấn Y Tế: Cung cấp thông tin cho bác sĩ và chuyên gia y tế để họ có thể tư vấn và hỗ trợ bệnh nhân hiệu quả.
- Nghiên cứu và Phân tích: Sử dụng kết quả để nghiên cứu thêm về mối liên quan giữa các yếu tố và tình trạng bệnh về giấc ngủ.

→ Đem lại giá trị lớn cho việc định hình chăm sóc sức khỏe và quản lý bệnh lý.

Tài liệu tham khảo

[1] Bài giảng học phần Khoa Học Dữ Liệu, "Orange Data Mining" khoa Công Nghệ Thông Tin Kinh Doanh, Đại học Kinh tế Tp.HCM, 2023.

[2] Bài giảng học phần Lập trình phân tích dữ liệu, khoa Công Nghệ Thông Tin Kinh Doanh, Đại học Kinh tế Tp.HCM, 2023.

[3] Bài giảng học phần Biểu diễn trực quan dữ liệu, "Data Visualization" khoa Công Nghệ Thông Tin Kinh Doanh, Đại học Kinh tế Tp.HCM, 2023.

[4] Bộ dữ liệu "Sleep Health and Lifestyle Dataset" trên Kaggle.

[5] Các phương pháp đánh giá mô hình máy học. - Quý blog -

[6] 5 WAYS TO HANDLE MISSING VALUES IN PYTHON by -Ibekwe kingsley-