BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC UEH
TRƯỜNG CÔNG NGHỆ và THIẾT KẾ
---o0o---
KHOA CÔNG NGHỆ THÔNG TIN KINH DOANH

ĐỒ ÁN KẾT THÚC HỌC PHẦN
MÔN BIỂU DIỄN TRỰC QUAN DỮ LIỆU

Chủ đề:

CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN TÌNH TRẠNG BÉO PHÌ

Giảng viên hướng dẫn: Ts. Nguyễn An Tế
Thực hiện: Nhóm 7
Danh sách nhóm: Lâm Thy Nhã
Võ Yến Nhi
Nguyễn Lê Thanh Oanh
Vương Kiến Phát
Lê Đình Phong

Lớp học phần: 4C1INF50904401
TP. Hồ Chí Minh, tháng 11 năm 2024

BẢNG ĐÁNH GIÁ THÀNH VIÊN

Họ và tên	Mã số sinh viên	Mức độ đóng góp
Nguyễn Lê Thanh Oanh	31221020761	100%
Lâm Thy Nhã	31211020484	100%
Vương Kiến Phát	31221022681	100%
Lê Đình Phong	31221025185	100%
Võ Yến Nhi	31221026992	100%

DANH MỤC BẢNG BIỂU

& HÌNH ẢNH

MỤC LỤC

Chương 1. Tổng quan đề tài

1. Sơ lược về đề tài

Báo Tuổi trẻ cho biết: *"Số người mắc bệnh béo phì đã tăng gấp 4 lần kể từ năm 1990. Theo WHO, béo phì thường phổ biến ở các nước nghèo và ngày càng nhiều trẻ em, thanh thiếu niên mắc bệnh hơn người lớn"*. Béo phì gây ra các vấn đề về thể chất và tinh thần, là vấn đề sức khỏe toàn cầu với những hậu quả nghiêm trọng. Đứng trước tình trạng tỷ lệ béo phì đang gia tăng đều đặn, nhóm chúng em lựa chọn sử dụng bộ dữ liệu Obesity Levels để làm nội dung báo cáo cho học phần Biểu diễn trực quan dữ liệu.

2. Mục tiêu nghiên cứu

- Trực quan hóa bằng các biểu đồ để xác định các yếu tố ảnh hưởng tới tình trạng béo phì bao gồm các đặc điểm các nhân (tuổi, giới tính, chiều cao, cân nặng), thói quen sinh hoạt (chế độ ăn uống, tần suất vận động, hút thuốc), hay tiền sử gia đình.
- Phân tích các biểu đồ trực quan nhằm khám phá mối quan hệ giữa các yếu tố trên với nguy cơ mắc bệnh béo phì.
- Hướng tới việc giúp đưa ra các khuyến nghị thực tế để cải thiện lối sống, giảm thiểu nguy cơ mắc bệnh béo phì và góp phần nâng cao sức khỏe cộng đồng.

3. Phương pháp nghiên cứu

Nhóm sử dụng nhiều phương pháp khác nhau theo nội dung nghiên cứu:

- Phân tích đơn biến sử dụng thống kê mô tả
- Phân tích đa biến sử dụng kiểm định Chi-square, One-way ANOVA.
- Các loại biểu đồ đặc thù nhằm mục đích phân tích chi tiết theo từng nội dung.

4. Tài nguyên sử dụng

- Bộ dữ liệu: **Obesity Levels** được thu thập từ Kaggle.
- Ngôn ngữ sử dụng : Python
- Môi trường làm việc: Jupyter Notebook

Chương 2. Tổng quan bộ dữ liệu

1. Tổng quan và các thuộc tính của bộ dữ liệu

Bộ dữ liệu Obesity Levels bao gồm dữ liệu để ước tính mức độ béo phì ở các cá nhân đến từ các quốc gia Mexico, Peru và Colombia, dựa trên thói quen ăn uống và tình trạng thể chất của họ.

Dữ liệu chứa 2111 dòng dữ liệu và 17 thuộc tính. Các bản ghi được dán nhãn với biến lớp NObesity (Mức độ béo phì), cho phép phân loại dữ liệu bằng cách sử dụng các giá trị Cân nặng không đủ, Cân nặng bình thường, Mức độ thừa cân I, Mức độ thừa cân II, Béo phì Loại I, Béo phì Loại II và Béo phì Loại III. 77% dữ liệu được tạo tổng hợp bằng công cụ Weka và bộ lọc SMOTE, 23% dữ liệu được thu thập trực tiếp từ người dùng thông qua nền tảng web.

Tên biến	Mô tả	Miền giá trị	Kiểu dữ liệu
Age	Tuổi của cá nhân	Số thực	Định lượng
Gender	Giới tính	Nam Nữ	Định danh
Height	Chiều cao	Số thực	Định lượng
Weight	Cân nặng (đơn vị: kg)	Số thực	Định lượng
CALC	Tần suất sử dụng đồ uống có chứa cồn	No (Không sử dụng) Sometimes (Đôi khi) Frequently (Thường xuyên) Always (Luôn luôn)	Định danh
FAVC	Thường xuyên sử dụng thực phẩm nhiều calo hay không?	Yes (Có) No (Không)	Định danh
FCVC	Tần suất tiêu thụ rau củ trong các bữa ăn (đơn vị: 100gram)	1 2 3	Định lượng
NCP	Số bữa ăn chính mỗi ngày (đơn vị: bữa)	1 2 3	Định lượng
SCC	Có theo dõi lượng calo mà người đó ăn hàng	Yes (Có) No (Không)	Định danh

	ngày hay không?		
SMOKE	Có hút thuốc không?	Yes (Có) No (Không)	Định danh
CH2O	Lượng nước uống mỗi ngày (đơn vị: L)	1 2 3	Định lượng
family_history_with_overweight	Tiền sử gia đình có người mắc béo phì không?	Yes (Có) No (Không)	Định danh
FAF	Tần suất hoạt động thể dục thể thao trong một tuần. (đơn vị: ngày)	0 1 2 3	Định lượng
TUE	Thời gian sử dụng các thiết bị điện tử mỗi ngày (đơn vị: giờ)	0 1 2	Định lượng
CAEC	Tần suất ăn vặt (các bữa ăn khác ngoài bữa chính)	No (Không) Sometimes (Thỉnh thoảng) Frequently (Thường xuyên) Always (Luôn luôn)	Định danh
MTRANS	Phương tiện thường sử dụng để di chuyển	Automobile (Xe ô tô) Motorbike (Xe máy) Bike (Xe đạp) Public Transportation (Phương tiện công cộng) Walking (Đi bộ)	Định danh
NObeyesdad	Mức độ béo phì	Insufficient weight (Thiếu cân) Normal Weight (Cân nặng bình thường) Overweight_Level_I (Thừa cân loại 1) Overweight_Level_II (Thừa cân loại 2) Obesity_Type_I	Định danh

		(Béo phì loại 1) Obesity_Type_II (Béo phì loại 2) Obesity_Type_III (Béo phì loại 3)	

Bảng 1: Bảng mô tả thuộc tính bộ dữ liệu gốc

2. Mô tả tổng quan bộ dữ liệu ban đầu

2.1. Kiểm tra dữ liệu:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2111 entries, 0 to 2110
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Age                             2111 non-null   float64
 1   Gender                          2111 non-null   object
 2   Height                          2111 non-null   float64
 3   Weight                          2111 non-null   float64
 4   CALC                            2111 non-null   object
 5   FAVC                            2111 non-null   object
 6   FCVC                            2111 non-null   float64
 7   NCP                             2111 non-null   float64
 8   SCC                             2111 non-null   object
 9   SMOKE                           2111 non-null   object
 10  CH2O                            2111 non-null   float64
 11  family_history_with_overweight  2111 non-null   object
 12  FAF                             2111 non-null   float64
 13  TUE                             2111 non-null   float64
 14  CAEC                            2111 non-null   object
 15  MTRANS                          2111 non-null   object
 16  NObeyesdad                      2111 non-null   object
dtypes: float64(8), object(9)
memory usage: 280.5+ KB
```

Hình 1: Kết quả trả ra của lệnh data.info()

Sử dụng câu lệnh data.info() để quan sát tổng quan bộ dữ liệu (tổng số dòng, số thuộc tính, kiểu dữ liệu từng cột, số giá trị null). Có thể thấy dữ liệu không có các giá trị missing values bởi vì tất cả các thuộc tính đều non-null.

2.2. Khám phá các biến định lượng:

2.2.1. Thống kê mô tả

	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
count	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000
mean	24.312600	1.701677	86.586058	2.419043	2.685628	2.008011	1.010298	0.657866
std	6.345968	0.093305	26.191172	0.533927	0.778039	0.612953	0.850592	0.608927
min	14.000000	1.450000	39.000000	1.000000	1.000000	1.000000	0.000000	0.000000
25%	19.947192	1.630000	65.473343	2.000000	2.658738	1.584812	0.124505	0.000000
50%	22.777890	1.700499	83.000000	2.385502	3.000000	2.000000	1.000000	0.625350
75%	26.000000	1.768464	107.430682	3.000000	3.000000	2.477420	1.666678	1.000000
max	61.000000	1.980000	173.000000	3.000000	4.000000	3.000000	3.000000	2.000000

Hình 2: Mô tả thống kê các biến định lượng

- ***count***: Số lượng giá trị không thiếu.
- ***mean***: Giá trị trung bình của mỗi cột.
- ***std***: Độ lệch chuẩn.
- ***min, 25%, 50%, 75%, max***: Giá trị nhỏ nhất, phân vị thứ 25, trung vị, phân vị thứ 75, và giá trị lớn nhất.

Từ bảng thống kê ta có thể thấy mẫu chủ yếu là người trẻ, với độ tuổi trung bình khoảng 24 và dao động từ 14 đến 61 tuổi. Chiều cao trung bình là 1,70m và cân nặng trung bình khá cao, khoảng 86,59kg, cho thấy sự đa dạng về vóc dáng trong nhóm. Chế độ ăn uống của họ tương đối cân bằng, với tần suất ăn rau trung bình 2,42 lần/ngày, số bữa ăn chính 2,69 bữa/ngày, và lượng nước tiêu thụ khoảng 2 lít/ngày. Tuy nhiên, hoạt động thể chất trung bình khá thấp, chỉ 1 giờ/tuần, cho thấy lối sống ít vận động. Thời gian sử dụng thiết bị điện tử trung bình là 0,66 giờ/ngày, không quá cao so với các nghiên cứu thường thấy. Nhìn chung, nhóm này phản ánh lối sống hiện đại với chế độ ăn uống tương đối ổn định nhưng cần cải thiện về mặt hoạt động thể chất để nâng cao sức khỏe.

	Gender	CALC	FAVC	SCC	SMOKE	family_history_with_overweight	CAEC	MTRANS	NObeyesdad
count	2111	2111	2111	2111	2111	2111	2111	2111	2111
unique	2	4	2	2	2	2	4	5	7
top	Male	Sometimes	yes	no	no	yes	Sometimes	Public_Transportation	Obesity_Type_I
freq	1068	1401	1866	2015	2067	1726	1765	1580	351

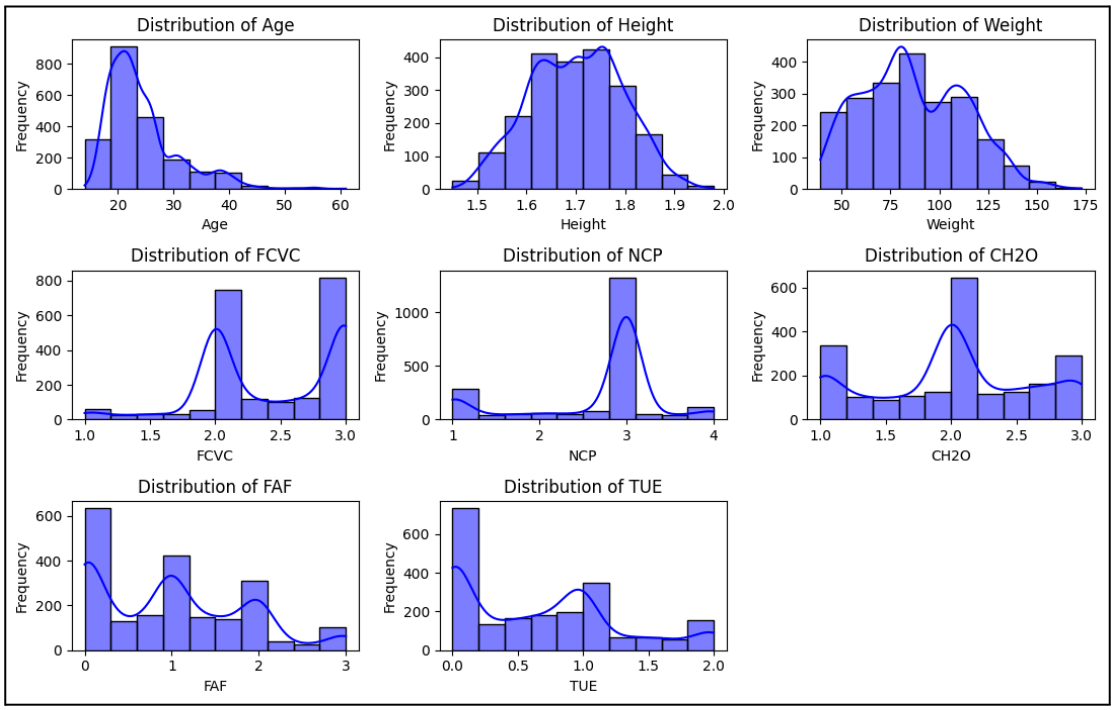
Hình 3: Mô tả thống kê các biến định tính

- ***count:*** Số lượng giá trị không thiếu.
- ***unique:*** Số lượng giá trị duy nhất trong cột.
- ***top:*** Giá trị xuất hiện nhiều nhất (mode).
- ***freq:*** Tần suất xuất hiện của giá trị phổ biến nhất.

Tập dữ liệu chiếm đa số là nam giới (1068 người), thói quen tiêu thụ đồ uống có cồn ở mức "Sometimes" (1401 người), và 1866 người thích đồ ăn nhanh. Phần lớn không hút thuốc (2067 người), nhưng 1726 người có tiền sử gia đình thừa cân. Tần suất ăn không kiểm soát phổ biến nhất là "Sometimes" (1765 người), và phương tiện di chuyển chính là

công cộng (1580 người). Loại béo phì phổ biến nhất là "Obesity_Type_I" (351 người). Nhóm mẫu chủ yếu có xu hướng thừa cân, thói quen ăn uống không lành mạnh (như tiêu thụ đồ ăn nhanh thường xuyên), và sử dụng phương tiện công cộng là chủ yếu.

2.2.2. Thống kê suy diễn



Hình 4: Các biểu đồ phân phối của các biến định lượng

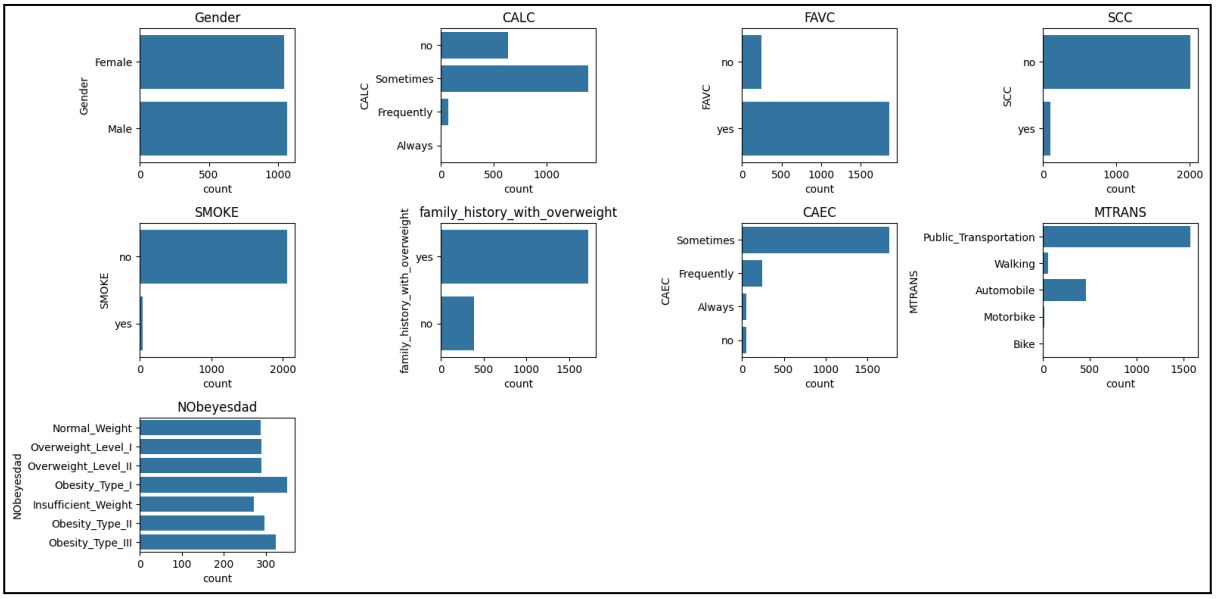
Dựa vào các biểu đồ phân phối của các biến định lượng ta thấy:

- ***Age***: Phân phối của tuổi cho thấy phần lớn các cá nhân nằm trong độ tuổi từ 20 đến 30, với một số ít ở độ tuổi cao hơn. Phân phối nghiêng trái, phản ánh rằng dữ liệu tập trung vào nhóm tuổi trẻ hơn.
- ***Height***: Chiều cao có phân phối gần chuẩn, tập trung chủ yếu từ 1.6 đến 1.8 mét. Các giá trị ngoài khoảng này khá hiếm.
- ***Weight***: Trọng lượng cho thấy một sự phân tán rộng, với đỉnh tập trung quanh mức 75–100 kg. Có một số cá nhân có trọng lượng rất cao.
- ***FCVC (Tần suất xuất hiện rau củ trong bữa ăn)***: Phân phối cho thấy phần lớn các cá nhân tiêu thụ rau củ ở mức 200 - 300gram.
- ***NCP (Số bữa ăn chính mỗi ngày)***: Phân phối rất tập trung, với hầu hết mọi người ăn 3 bữa mỗi ngày. Rất ít người ăn ít hơn hoặc nhiều hơn 3 bữa.
- ***CH2O (Lượng nước uống mỗi ngày):*** Phân phối đa đỉnh, đa số tập trung ở mức 2 lít nước mỗi ngày nhưng vẫn có một lượng lớn người uống 1 lít và 3 lít nước mỗi ngày.
- ***FAF (Tần suất hoạt động thể dục thể thao)***: Số giờ hoạt động thể chất có phân phối đa đỉnh, với một số lượng lớn cá nhân không tham gia hoặc tham gia ít, và số ít tham gia luyện tập thường xuyên hơn.

- ***TUE (Thời gian sử dụng thiết bị điện tử mỗi ngày)***: Biểu đồ phân phối có đỉnh tập trung ở mức thấp (dưới 1 giờ), phản ánh rằng phần lớn mọi người chỉ sử dụng thiết bị trong thời gian ngắn.

Nhận thấy rằng phân phối của các biến "FCVC", "NCP", "CH2O", "FAF", "TUE" chứa các giá trị nằm ngoài miền giá trị cho phép, nhóm cần phải xử lý đưa các giá trị này về dạng số nguyên.

2.3. Khám phá các biến định tính



Hình 5: Các biểu đồ tỷ lệ của các biến định tính

Ngoại trừ biến Gender và NObeyesdad có phân phối khá đồng đều thì các biến còn lại đều có tần số bị thiên lệch về một hoặc hai giá trị nhất định trong miền giá trị.

Chương 3. Tiền xử lý dữ liệu

1. Kiểm tra định dạng dữ liệu

1.1. Kiểu tra biến có dữ liệu số nguyên

```
print(isinstance(df['FCVC'][0], int))
print(isinstance(df['TUE'][0], int))
print(isinstance(df['NCP'][0], int))
print(isinstance(df['CH2O'][0], int))
print(isinstance(df['FAF'][0], int))
```

```
False
False
False
False
False
```

5 biến “FCVC”, “NCP”, “CH2O”, “FAF”, “TUE” không chứa giá trị integer. Nhóm tiến hành chuyển đổi từ ‘float64’ sang ‘int’ bằng cách làm tròn giá trị đến phần nguyên

```
df["TUE"] = df["TUE"].round(0).astype(int)
df["FCVC"] = df["FCVC"].round(0).astype(int)
df["NCP"] = df["NCP"].round(0).astype(int)
df["CH2O"] = df["CH2O"].round(0).astype(int)
df["FAF"] = df["FAF"].round(0).astype(int)
```

1.2. Kiểm tra biến nhị phân

```
binary_variables = ['family_history_with_overweight', 'FAVC', 'SMOKE']
variables_not_binary = []

for var in binary_variables:
    if len(df[var].unique()) != 2:
        variables_not_binary.append(var)

print("Danh sách các biến không có 2 miền giá trị:", variables_not_binary)
```

```
Danh sách các biến không có 2 miền giá trị: []
```

1.3. Kiểm tra định dạng các biến định tính:

```
# Kiểm tra kiểu dữ liệu của các biến và in ra các biến không phải categorical
categorical_cols = ['Gender', 'CAEC', 'CALC', 'MTRANS', 'NObeyesdad']
```

```
# Danh sách lưu trữ các biến không phải categorical
non_categorical = []

for col in categorical_cols:
    if not (df[col].dtype == 'object' or pd.api.types.is_categorical_dtype(df[col])):
        non_categorical.append(col)

# In ra các biến không phải categorical
if non_categorical:
    print("Các biến KHÔNG phải là categorical:", non_categorical)
else:
    print("Tất cả các biến đều là categorical.")
```

```
Tất cả các biến đều là categorical.
```

1.4. Kiểm tra định dạng các biến liên tục:

```
continuous_cols = ['Age', 'Height', 'Weight', 'NCP', 'FAF', 'CH2O']

# Danh sách lưu trữ các biến không liên tục
non_continuous = []

# Kiểm tra kiểu dữ liệu và tính liên tục
for col in continuous_cols:
    if not (pd.api.types.is_numeric_dtype(df[col]) and df[col].nunique() > 10):
        non_continuous.append(col)

# In ra các biến không phải liên tục
if non_continuous:
    print("Các biến KHÔNG phải là liên tục:", non_continuous)
else:
    print("Tất cả các biến đều là liên tục.")
```

Tất cả các biến đều là liên tục.

2. Thay đổi tên các biến

Hiện tại trong bộ dữ liệu, các tên biến đang chưa thể hiện rõ được toàn bộ ý nghĩa của nó và cách đặt tên vẫn chưa được thống nhất. Vì vậy nhóm quyết định thay đổi tên các biến như sau:

```
#Thay đổi tên của các cột
name_map = {
    "Age": "age",
    "Gender": "gender",
    "Height": "height",
    "Weight": "weight",
    "CALC": "alcohol",
    "MTRANS": "transportation",
    "NObeyesdad": "obesity",
    "FAVC": "high caloric",
    "FCVC": "vegetables",
    "NCP": "main meals",
    "SCC": "cal monitor",
    "SMOKE": "smoke",
    "CH2O": "water",
    "FAF": "physical activity",
    "TUE": "tech devices",
    "CAEC": "between meals",
    "family_history_with_overweight": "family history"
}
```

3. Kiểm tra và xử lý Missing Values

Để kiểm tra xem những cột nào có chứa missing values, nhóm thực hiện tính tính tổng số missing values ở mỗi cột bằng phương thức isnull().sum()

```
# Kiểm tra giá trị trống
missing_values = df.isnull().sum()
```

```
print('Các giá trị bị thiếu:')
print(missing_values)
```

```
Các giá trị bị thiếu:
age                 0
gender              0
height              0
weight              0
alcohol             0
high caloric        0
vegetables          0
main meals          0
cal monitor         0
smoke               0
water               0
family history      0
physical activity   0
tech devices        0
between meals       0
transportation      0
obesity             0
```

Kết quả trả về tất cả các cột đều không tồn tại giá trị missing values nào nên không cần bước xử lý missing values.

4. Kiểm tra các giá trị bất thường của biến định lượng

Đối với các biến định lượng trong bộ dữ liệu này, việc xuất hiện giá trị âm là một điều bất thường. Do đó, nhóm tiến hành kiểm tra sự tồn tại của giá trị bất thường trong các biến định lượng

```
# Kiểm tra các giá trị bất thường
abnormal_values = pd.DataFrame()
for column in df.columns:
    if df[column].dtype != 'object':  # Kiểm tra nếu cột là số học
        abnormal = df[df[column] < 0]  # Kiểm tra giá trị âm
        abnormal_values = pd.concat([abnormal_values, abnormal])

print("Các giá trị bất thường:")
print(abnormal_values)
```

```
Các giá trị bất thường:
Empty DataFrame
Columns: [age, gender, height, weight, alcohol, high caloric, vegetables, main meals, cal monitor, smoke, water, family history, physical activity, tech devices, between meals, transportation, obesity]
```

```
Index: []
```

Kết quả là không có giá trị bất thường.

5. Kiểm tra và xử lý giá trị trùng lặp

Nhóm tiến hành kiểm tra các dòng bị trùng lặp và loại bỏ nó ra khỏi bộ dữ liệu

```
# Count duplicated rows
num_duplicates = df.duplicated().sum()
print(f"Number of duplicated samples: {num_duplicates}")

# Drop duplicated rows and keep the first occurrence
df_no_duplicates = df.drop_duplicates(keep='first')
```

```
Number of duplicated samples: 24
```

6. Thêm thuộc tính BMI

BMI (Body Mass Index - Chỉ số khối cơ thể) là một chỉ số dùng để đánh giá mối quan hệ giữa cân nặng và chiều cao của một người, qua đó xác định tình trạng cơ thể thuộc nhóm gầy, bình thường, thừa cân, hay béo phì. Chỉ số BMI được tính bằng cách lấy cân nặng của cơ thể tính bằng kilogram và chia cho bình phương của chiều cao tính bằng mét. Việc thêm biến BMI cung cấp thông tin tổng quát về tình trạng cơ thể (thiếu cân, bình thường, thừa cân, hoặc béo phì), thay vì chỉ nhìn vào cân nặng hoặc chiều cao riêng lẻ. Có thể coi biến BMI coi là một đặc trưng bổ sung, giúp cải thiện hiệu suất của các mô hình dự đoán.

Chỉ số BMI thường được dùng để phân loại béo phì theo các mức độ như sau

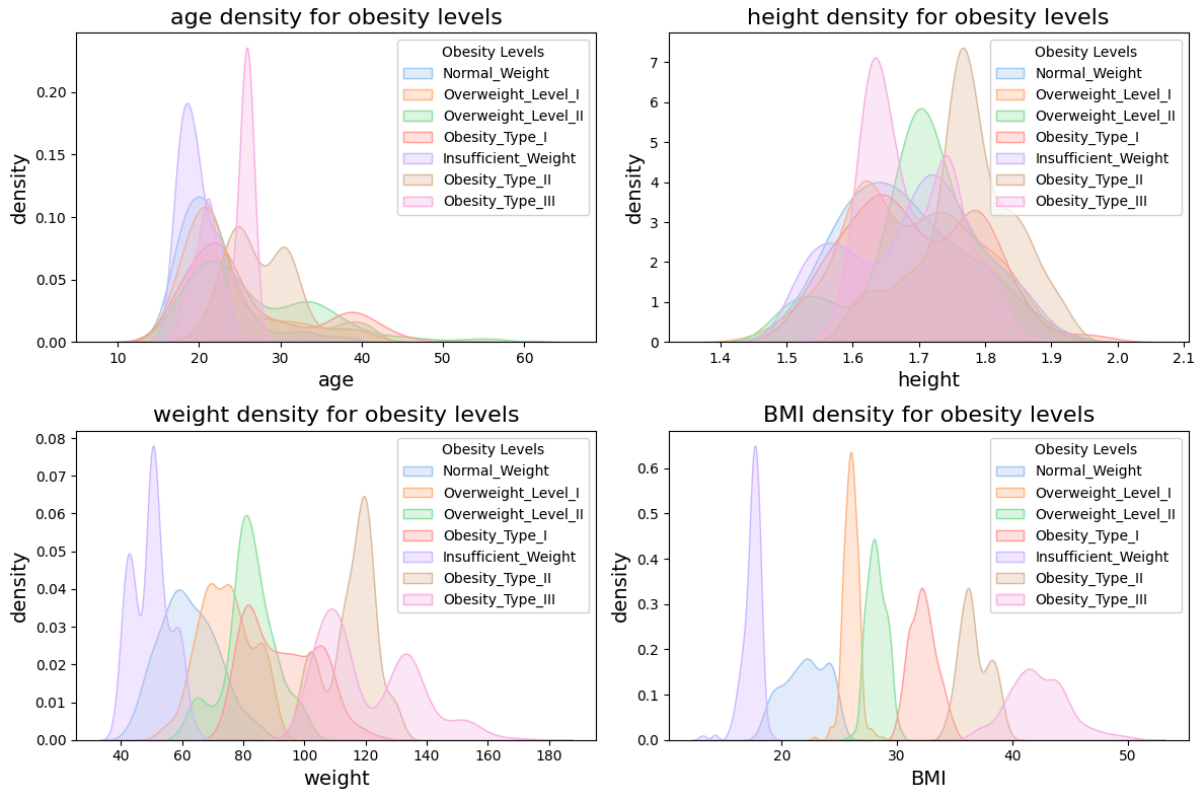
Bảng đánh giá theo chuẩn của Tổ chức Y tế thế giới(WHO) và dành riêng cho người châu Á (IDI&WPRO):		
Phân loại	**WHO BMI (kg/m2)**	**IDI & WPRO BMI (kg/m2)**
Cân nặng thấp (gầy)	<18.5	<18.5
Bình thường	18.5 - 24.9	18.5 - 22.9
Thừa cân	25 - 29.9	23 - 24.9
Béo phì độ I	30 - 34.9	25 - 29.9
Béo phì độ II	35 - 39.9	30 - 34.9
Béo phì độ III	>=40	>=35

Nguồn: Trường Cao đẳng Y khoa Phạm Ngọc Thạch

```
df['BMI'] = round(df['weight'] / (df['height']) ** 2, 2)
```

7. Biểu đồ thể hiện sự tương quan giữa các biến độc lập với biến phụ thuộc:

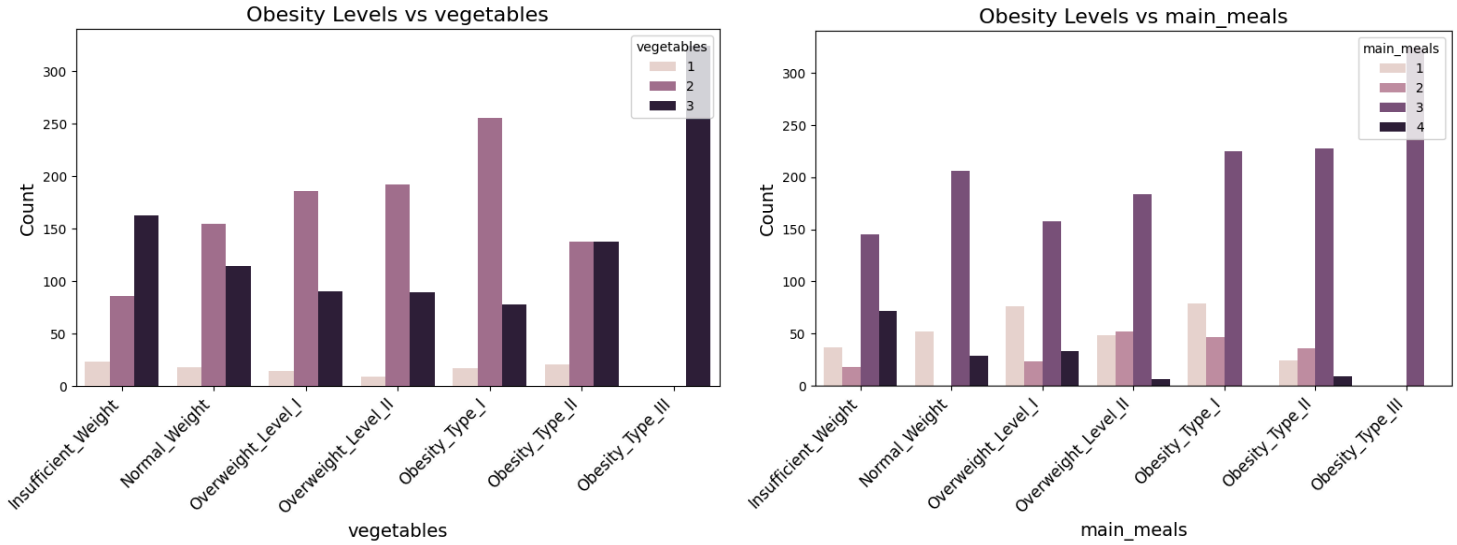
7.1. Biến liên tục:



Hình 6: Mật độ của các biến liên tục so với biến phụ thuộc

Phân phối của biến BMI và cân nặng so với nhóm tuổi cho thấy sự phân tách rõ ràng, đặc biệt là biến BMI, thể hiện sự tương quan cao với biến phụ thuộc. Trong phân phối của biến này, ta thấy có một số khoảng mà trong đó có nhiều hơn một mức độ béo phì. Điều này thể hiện rằng cơ thể mỗi người có các thành phần với những định lượng khác nhau, do đó có những người có cùng cân nặng chiều cao nhưng mức độ béo phì của họ lại khác nhau.

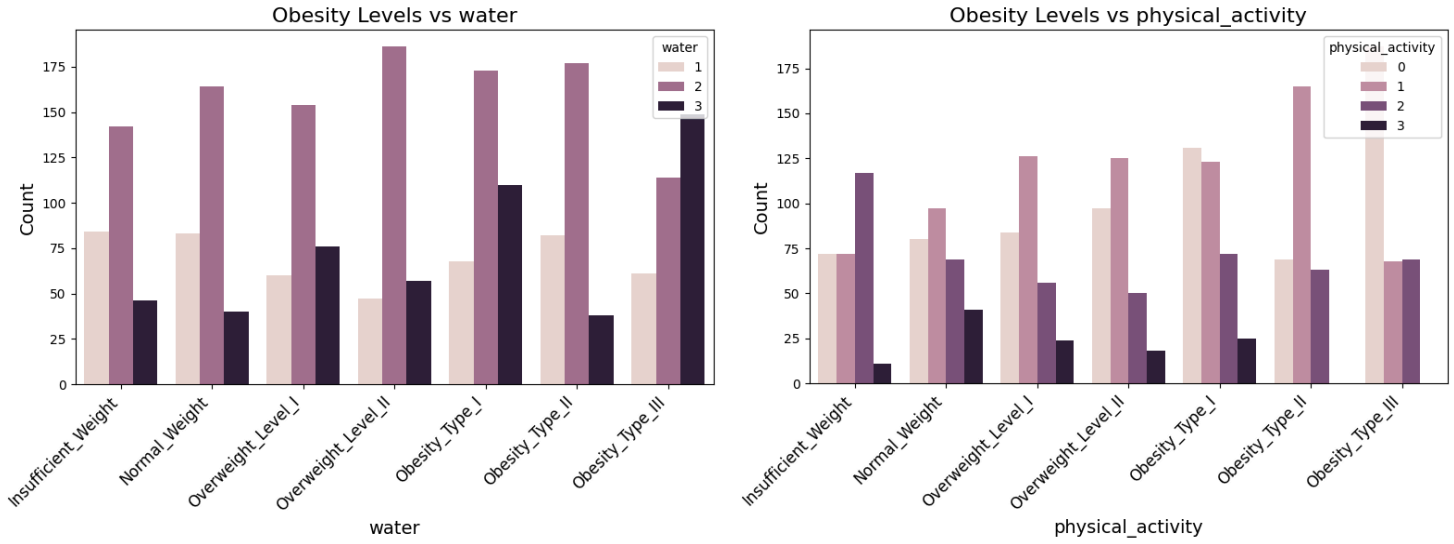
7.2. Biến rời rạc



Hình 7: Phân phối của biến vegetables và biến main_meals so với biến phụ thuộc

Lượng rau củ ăn trong mỗi bữa không tương quan cao với mức độ béo phì, vì những người ăn trên 200 gram rau củ vẫn chiếm đa số là người thừa cân và béo phì, tuy nhiên nhóm này có rất ít người thuộc diện béo phì loại 3. Nhóm người ăn 300 gram rau củ một ngày lại có nhiều người thuộc nhóm thiếu cân và béo phì loại 3, trong khi tần số ít hơn ở nhóm từ cân nặng bình thường đến béo phì loại 2.

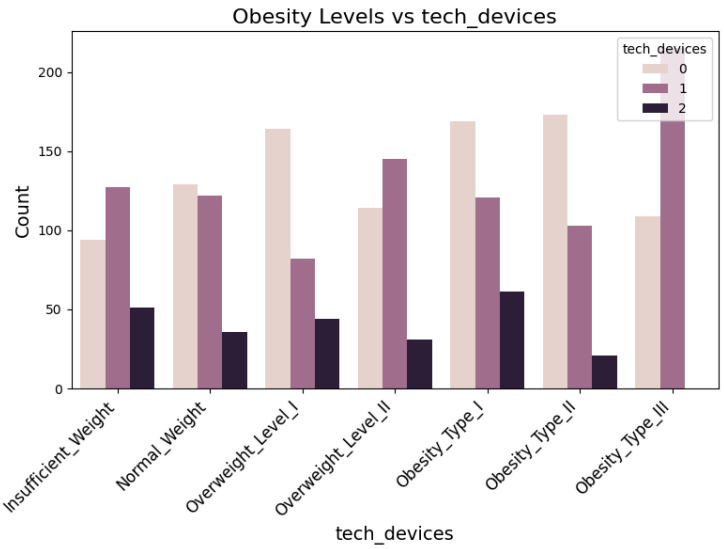
Trong số những người ăn từ 1 đến 2 bữa một ngày có rất người thuộc nhóm béo phì loại 3. Điều thú vị là những người ăn 4 bữa một ngày lại có tần suất thuộc nhóm thiếu cân cao hơn hẳn các nhóm còn lại. Còn những người ăn 3 bữa một ngày lại có tần suất thuộc nhóm béo phì loại 3 cao nhất, các mức béo phì loại 1, 2 và cân nặng bình thường cũng có tần suất cao và gần như là tương đương.



Hình 8: Phân phối của biến water và biến physical_activity so với biến phụ thuộc

Nhóm người uống 1 lít nước mỗi ngày mặc dù có phân phối tần suất của mức độ béo phì khá đều, trong đó nhóm thiếu cân và cân nặng bình thường chiếm tần suất lớn nhất. Nhóm người uống 2 lít nước mỗi ngày lại có tần suất người thuộc diện béo phì loại 3 thấp nhất, còn lại các mức độ béo phì khác có phân phối khá đồng đều. Nhóm người uống 3 lít nước một ngày lại là nhóm có tần suất người béo phì loại 3 cao nhất trong cả 3 nhóm và cao ở các nhóm thừa cân, béo phì so với nhóm thiếu cân và cân nặng bình thường.

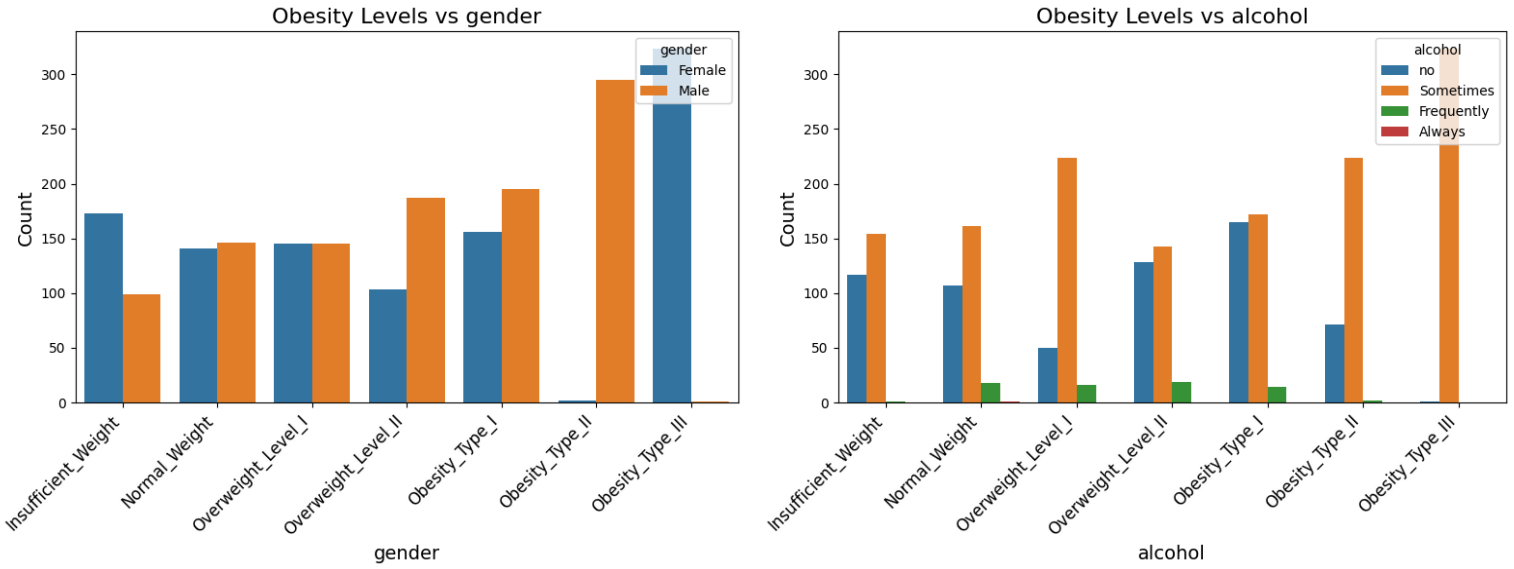
Nhóm người không tập thể dục có tần suất mắc thừa cân béo phì cao, cao nhất là béo phì loại 3. Nhóm người tập thể dục 1 lần/tuần cũng có tần suất thừa cân béo phì cao, nhưng loại 2 là cao nhất. Nhóm người tập thể dục 2 lần/tuần lại có tần suất của các mức độ béo phì khá đồng đều, tuy nhiên lại cao ở nhóm thiếu cân. Nhóm người tập thể dục 3 lần/ tuần có tần số cân nặng bình thường chiếm tỉ lệ cao nhất, tần số của các nhóm thừa cân và béo phì loại một cũng khá cao, nhưng gần như không có béo phì loại 2 và loại 3.



Hình 9: Phân phối của biến tech_devices so với biến phụ thuộc

Nhóm người sử dụng điện thoại 2 tiếng một ngày không có quan sát thuộc nhóm béo phì loại 3. Nhóm người sử dụng điện thoại 1 tiếng một ngày có tần suất mắc béo phì loại 3 cao hơn hẳn các mức độ béo phì còn lại. Những người không sử dụng điện thoại vẫn có tần suất mắc béo phì cao.

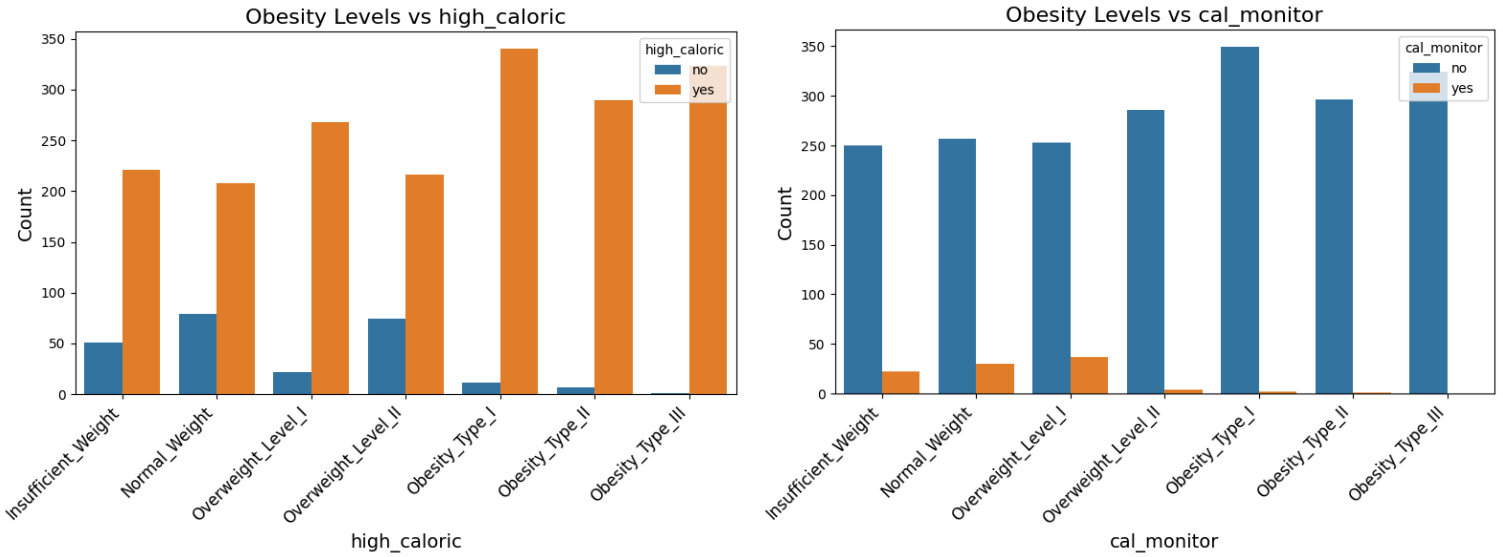
7.3. Biến định tính



Hình 10: Phân phối của biến gender và biến alcohol so với biến phụ thuộc

Ở nữ giới gần như không xuất hiện giá trị béo phì loại 2 và mức độ béo phì loại 3 chiếm tần suất lớn nhất. Còn ở nam giới không có giá trị béo phì loại 3 và béo phì mức độ 2 có tần suất xuất hiện nhiều nhất.

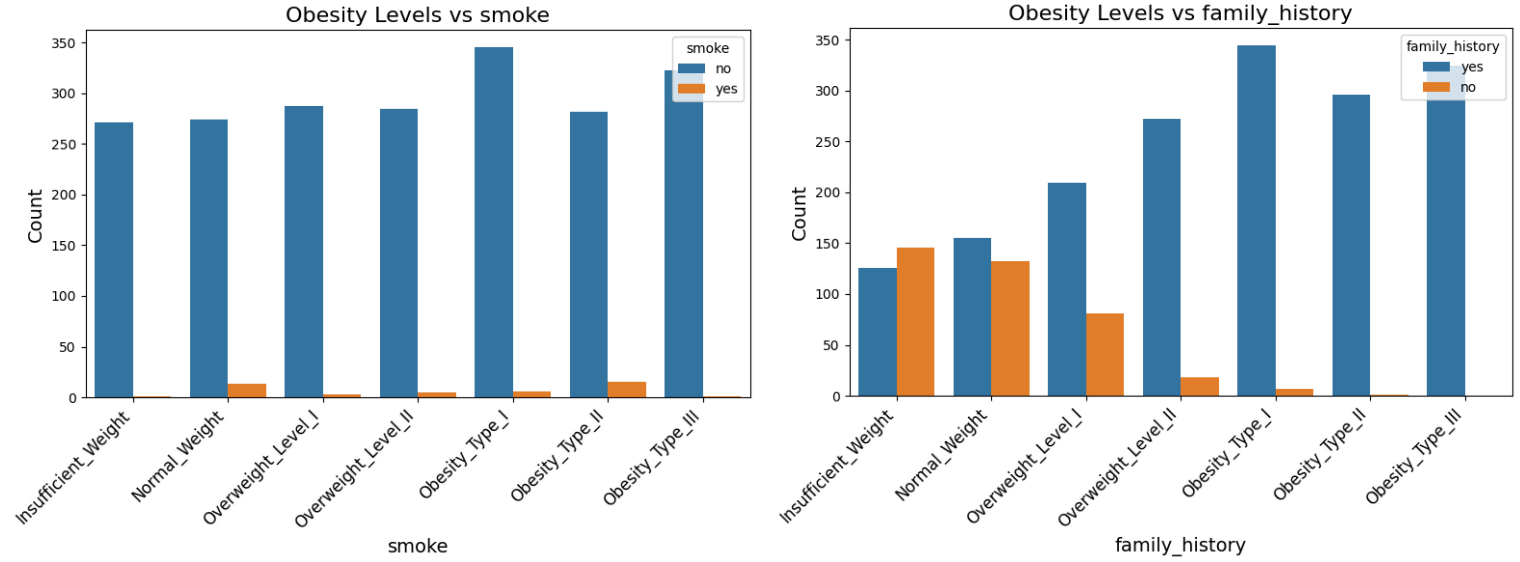
Không có nhiều sự tương quan giữa mức độ béo phì và tần suất sử dụng sản phẩm có cồn. Cả nhóm người thường xuyên sử dụng và không sử dụng đồ uống có cồn đều ít rơi vào nhóm béo phì loại 2 loại 3. Trong khi đó nhóm người thỉnh thoảng sử dụng lại có số lượng người thuộc nhóm béo phì loại 3 cao.



Hình 11: Phân phối của biến high_caloric và biến cal_monitor so với biến phụ thuộc

Những người không tiêu thụ thực phẩm có nhiều calories thường xuyên ít rơi vào các nhóm béo phì, ngược lại những người thường xuyên tiêu thụ thực phẩm có nhiều calories lại rơi vào nhóm mức độ béo phì nhiều hơn so với các mức độ còn lại.

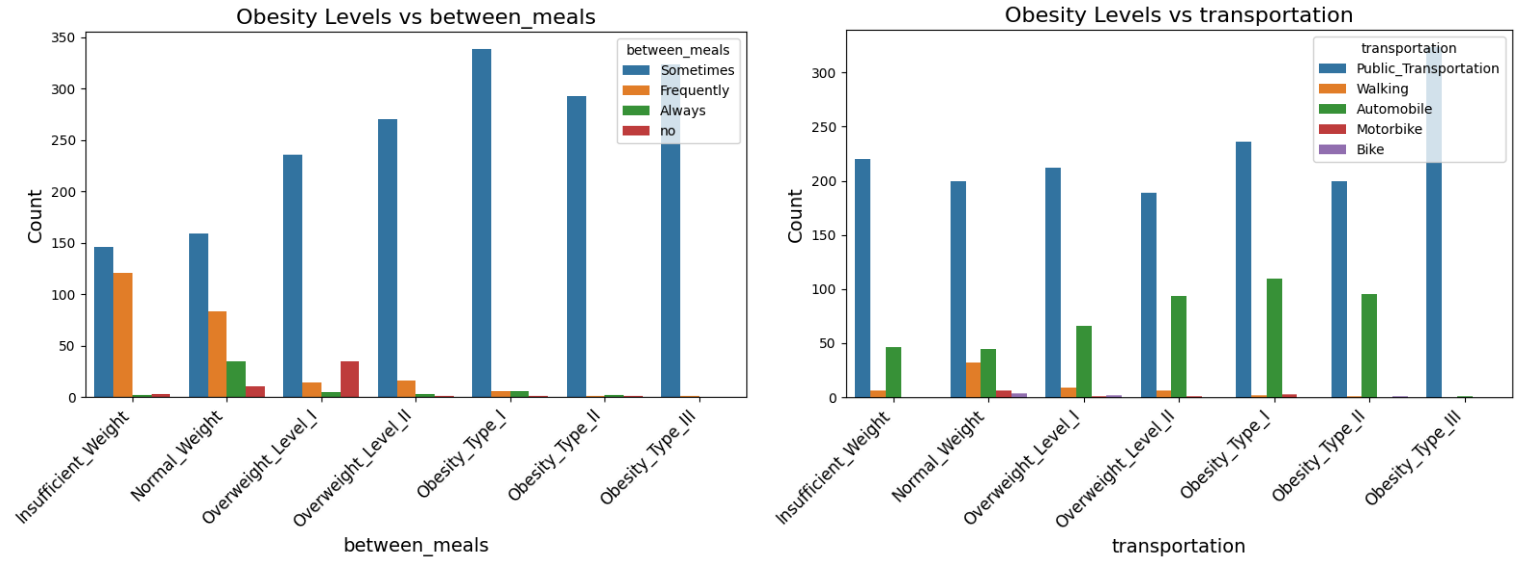
Đa số những người có kiểm soát lượng calories nạp vào cơ thể rơi vào mức từ thiếu cân đến thừa cân loại 1. Những người không kiểm soát lượng calories có mức độ béo phì rơi nhiều vào nhóm béo phì từ loại 1 đến loại 3.



Hình 12: Phân phối của biến smoke và biến family_history so với biến phụ thuộc

Việc có hút thuốc hay không không cho thấy nhiều sự ảnh hưởng đến mức độ béo phì, vì phân phối về mức độ béo phì của những người hút thuốc và không hút thuốc khá đồng đều,

Những người không có tiền sử gia đình mắc béo phì chủ yếu rơi vào các nhóm thiếu cân, cân nặng bình thường và thừa cân loại 1, trong khi đó những người có tiền sử gia đình mắc béo phì rơi nhiều vào nhóm thừa cân và béo phì.



Hình 13: Phân phối của biến between_meals và biến transportation so với biến phụ thuộc

Những người có thường xuyên ăn các bữa phụ giữa các bữa chính lại thuộc nhiều vào nhóm thiếu cân hoặc cân nặng bình thường, trong khi thừa cân loại 1 lại chiếm đa số trong số những người không ăn bữa phụ. Đa số những người thỉnh thoảng ăn bữa phụ là người thừa cân cho đến béo phì.

Những người chọn đi bộ, đi xe đạp và xe máy phần lớn có cân nặng bình thường, trong khi đó những người sử dụng phương tiện công cộng hoặc đi xe ô tô phần nhiều là người thừa cân và béo phì. Điều thú vị là người đi ô tô lại gần như không có thừa cân loại nặng 3.

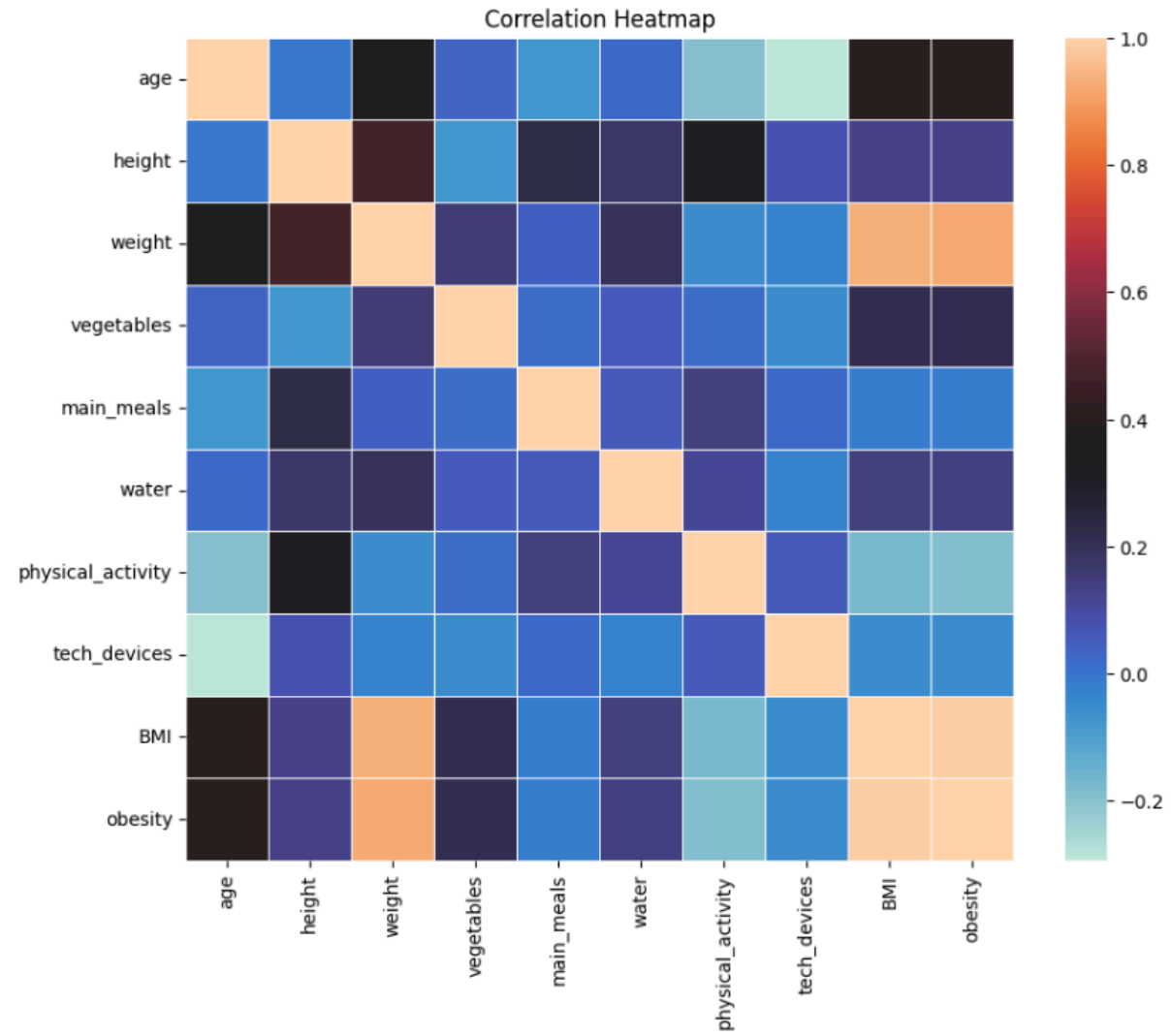
Chương 4. Kiểm định giả thuyết

1. Ma trận tương quan

Đầu tiên, nhóm thực hiện tính toán ma trận tương quan giữa các biến trong bộ dữ liệu đã qua tiền xử lý sử dụng phương pháp Spearman. Biến mx sẽ cho kết quả là một ma trận với các giá trị tương quan từ -1 đến 1.

```
mx = f_df.corr(method='spearman')
plt.figure(figsize=(10, 8))
sns.heatmap(mx, cmap="icefire", linewidths=0.5)
plt.title("Correlation Heatmap")
plt.show()
```

Tiếp theo, nhóm sẽ trực quan hóa ma trận mx bằng cách vẽ biểu đồ nhiệt như hình bên dưới. Thông qua biểu đồ nhiệt, có thể nhận thấy rằng những cặp biến có màu sắc thuộc dãy màu đỏ cam sẽ tương quan với nhau rất mạnh. Ví dụ như là weight và BMI, weight và obesity, BMI và obesity.



Hình 14: Ma trận tương quan

Các cặp biến độc lập tương quan mạnh với nhau thường mang thông tin tương tự nhau, có thể gây ra hiện tượng đa cộng tuyến, dẫn đến việc các kiểm định giả thuyết sẽ không cần đáng tin cậy. Do đó, nhóm cần phải xác định các cặp biến độc lập bị tương quan quá mạnh với nhau và cân nhắc có nên loại bỏ một trong hai để tránh hiện tượng đa cộng tuyến hay không.

Nhóm xác định các cặp biến có hệ số tương quan ($|r| = 0.7$) sẽ là các cặp biến có tương quan mạnh với nhau. Đoạn mã bên dưới sẽ duyệt qua quan ma trận mx, lấy giá trị tương quan (correlation_value) ở mỗi ô (i, j). Nếu hệ số tương quan ở ô đó lớn hơn 0.7 hoặc bé hơn -0.7 thì cặp biến đó sẽ được ghi nhận là cặp biến có tương quan mạnh với nhau.

```
correlation_pairs = []
```

```
for i in range(len(mx.columns)):
    for j in range(i + 1, len(mx.columns)):
        col1 = mx.columns[i]
        col2 = mx.columns[j]
        correlation_value = mx.iloc[i, j]
        if abs(correlation_value) > 0.7:
            correlation_pairs.append((col1, col2, correlation_value))

# Hiển thị các cặp cột có tương quan lớn hơn 0.7
for col1, col2, value in correlation_pairs:
    print(f"Cặp features: {col1} và {col2}, Hệ số tương quan: {value:.2f}")
```

Kết quả trả về cũng tương tự như nhận xét các cặp biến tương quan ở biểu đồ nhiệt (Hình 1). Với các cặp biến weight và obesity ($r = 0.92$), BMI ($r = 0.99$) và obesity có tương quan mạnh với nhau và obesity là biến phụ thuộc nên 2 biến weight, BMI có ý nghĩa thống kê rất lớn với biến obesity. Còn cặp BMI và weight là 2 biến độc lặp nhưng chúng có $r = 0.93$, rất lớn nên 2 biến này có thể gây ra hiện tượng đa cộng tuyến.

```
Cặp features: weight và BMI, Hệ số tương quan: 0.93
Cặp features: weight và obesity, Hệ số tương quan: 0.92
Cặp features: BMI và obesity, Hệ số tương quan: 0.99
```

Hình 15: Kết quả các cặp biến có hệ số tương quan trên |0.7|

Sau đó, nhóm sử dụng F-Test để xác định các biến độc lập quan trọng, có ảnh hưởng mạnh với biến mục tiêu obesity. Đoạn mã sử dụng SelectKBest với việc chọn ra k = 9 đặc trưng tốt nhất của bộ dữ liệu, và sử dụng score_func=f_classif là F-test để đánh giá mức độ tương quan giữ từng biến độc lập X với biến mục tiêu Obesity.

fit_transform(X, y) để tính toán và chọn ra các đặc trưng tốt nhất từ X, kết quả trả về tập dữ liệu mới chỉ chứa các đặc trưng được chọn. k_best.get_support(indices=True) để lấy danh sách chỉ số của các đặc trưng được chọn. Và k_best.scores_ để trả về giá trị F của tất cả các đặc trưng, cho biết mức độ liên quan với biến mục tiêu. Và sau đó, in ra tên biến và kèm theo giá trị F của từng biến với top 10 biến có ảnh hưởng cao nhất với biến mục tiêu.

```
k_best = SelectKBest(score_func=f_classif, k=9) #dùng F-test đánh giá mức độ tương quan của mỗi feature
```

```
với target

X_ = k_best.fit_transform(X, y)
y_ = y

#Trả về danh sách chỉ số của các đặc trưng được chọn
selected_features_indices = k_best.get_support(indices=True)

#Trả về một mảng chứa giá trị F của tất cả các đặc trưng
feature_scores = k_best.scores_

#Tạo danh sách các tuples gồm các đặc trưng và score tương ứng
feature_info = list(zip(X.columns, feature_scores))

#Sắp xếp danh sách feature_info theo giá trị F giảm dần.
sorted_feature_info = sorted(feature_info, key=lambda x: x[1], reverse=True)
# in ra top 10 đặc trưng có giá trị F cao nhất
for feature_name, feature_score in sorted_feature_info[:10]:
    print(f"{feature_name}: {feature_score:.2f}")
```

Cũng như kết quả từ ma trận tương quan, BMI và weight là 2 biến có mối quan hệ mạnh và rất quan trọng với biến mục tiêu Obesity khi F_score hoàn toàn vượt trội so với các biến còn lại, lần lượt là 10,085.23 đối với biến BMI và 1,966.52 đối với biến weight. Biến vegetables và age dù thấp hơn nhiều so với 2 biến trên, nhưng cũng ảnh hưởng đến biến mục tiêu Obesity khi F-score của 2 biến đều trên 50. Còn biến tech_devices có ảnh hưởng khá nhỏ khi F_score chỉ bằng 7.69.

```
BMI: 10085.23
weight: 1966.52
vegetables: 91.02
age: 77.95
height: 38.43
main_meals: 25.30
water: 17.66
physical_activity: 17.20
tech_devices: 7.69
```

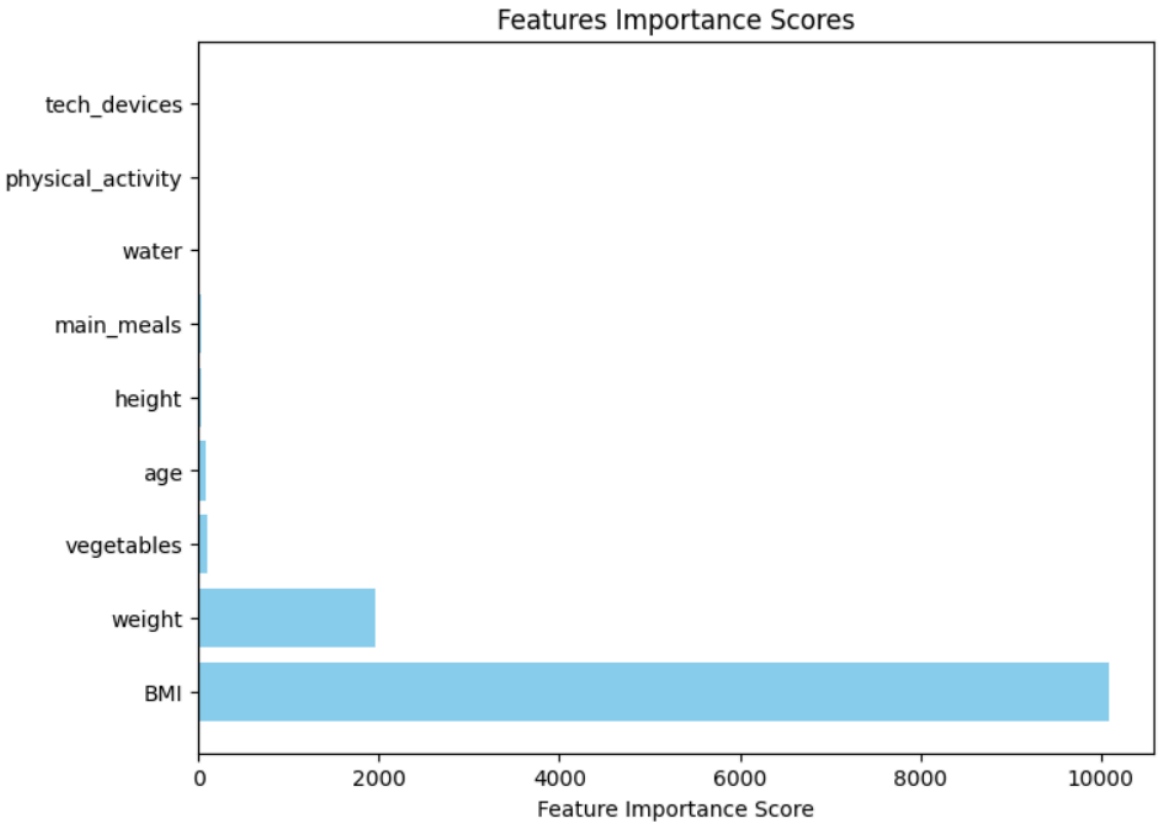
Hình 16: 10 biến độc lập có ảnh hưởng mạnh đến biến mục tiêu Obesity

Sau đó, nhóm sẽ thực hiện vẽ biểu đồ Bar Chart để trực quan hóa mức độ quan trọng của các đặc trưng đối với biến mục tiêu, với cột x là thể hiển F_score và cột y là các top 10 đặc trưng quan trọng đối với biến Obesity.

```
feature_names, feature_scores = zip(*sorted_feature_info[:])


plt.figure(figsize=(8, 6))
plt.barh(feature_names, feature_scores, color="skyblue")
plt.xlabel("Feature Importance Score")
plt.title("Features Importance Scores")
plt.show()
```

Như đã nhận xét bên trên, 2 đặc trưng weight và BMI có F_score rất cao nên nó rất quan trọng với biến Obesity. Biến BMI thì có F_score gần như vượt trội so với các biến còn lại. Biến age và vegetable thì cột hiển thị khá thấp nhưng vẫn có thể nhìn thấy được. Các biến còn lại thì hoàn toàn không thể nhìn thấy được cột F_score.



Hình 17: Biểu đồ cột thể hiện mức độ quan trọng của biến độc lập với biến Obesity

2. Kiểm định các biến định lượng với biến phụ thuộc

Đến với phần kiểm định, nhóm sẽ bắt đầu kiểm định với các thuộc tính định lượng trước, bao gồm các biến là ['age', 'vegetables', 'height', 'main meals', 'water', 'physical activity', 'tech devices']. Nhóm không thực hiện kiểm định với biến weight vì ở phần trước, nhóm đã đưa ra kết luận biến weight rất quan trọng với biến mục tiêu Obesity. Vì các biến được liệt kê trên không thể hiện được ảnh hưởng quá rõ ràng nên nhóm sẽ đi kiểm định chúng có thật sự quan trọng đối với biến Obesity hay không.

Nhóm sẽ sử dụng kiểm định ANOVA giữa các biến độc lập và biến mục tiêu. Các giả thuyết được đặt ra sẽ như sau:

- H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$: Trung bình của tất cả các nhóm là như nhau (với k là số nhóm)
- H_1: Tồn tại ít nhất một nhóm i sao cho $\mu_i \neq \mu$.

Kiểm định 1: *Độ tuổi (age) không ảnh hưởng đến Obesity*

- H_0: Trung bình độ tuổi là như nhau giữa các mức độ béo phì.
- H_1: Có sự khác biệt về trung bình độ tuổi giữa các mức độ béo phì với nhau.

Đoạn mã sẽ phân tách dữ liệu thành 7 nhóm dựa trên giá trị của biến obesity (mức độ béo phì). Mỗi nhóm chứa độ tuổi (age) của các cá nhân thuộc mức độ béo phì tương ứng:

- 1: Thiếu cân.
- 2: Cân nặng bình thường.
- 3, 4: Thừa cân cấp độ I và II.
- 5, 6, 7: Béo phì cấp độ I, II và III.

```
a_Obesity_Type_I = df[df['obesity'] == 5]['age']
a_Obesity_Type_II = df[df['obesity'] == 6]['age']
a_Obesity_Type_III = df[df['obesity'] == 7]['age']
a_Normal_Weight = df[df['obesity'] == 2]['age']
a_Overweight_Level_I = df[df['obesity'] == 3]['age']
a_Overweight_Level_II = df[df['obesity'] == 4]['age']
a_Insufficient_Weight = df[df['obesity'] == 1]['age']
```

Thực hiện kiểm định ANOVA bằng đoạn mã bên dưới cho biến age và các biến định lượng còn lại:

- Sử dụng hàm stats.f_oneway từ thư viện scipy.stats để tính giá trị F và trị số p.
- Tham số đầu vào là các nhóm giá trị age theo từng mức độ béo phì.

Sau đó, nhóm sẽ đi so sánh giá trị p với mức ý nghĩa $\alpha = 0.05$:

- Nếu $p < 0.05$: Bác bỏ H_0, kết luận có sự khác biệt về độ tuổi trung bình giữa các nhóm.
- Nếu $p \geq 0.05$: Không bác bỏ H_0, không tìm thấy sự khác biệt ý nghĩa thống kê.

```
f, p = stats.f_oneway(a_Obesity_Type_I,a_Obesity_Type_II,a_Obesity_Type_III,
a_Normal_Weight,a_Overweight_Level_I, a_Overweight_Level_II, a_Insufficient_Weight )
print(f"ANOVA Test: F-statistic = {f:.2f}, p-value = {p:.3f}")



alpha = 0.05
## Kết luận theo phương pháp p-value (trị số p)
if (p < alpha):
    print(f'Trị số p = {p:} < {alpha:.4f} cho nên bác bỏ H0 ==> Có sự khác biệt giữa trung bình độ tuổi và các mức độ béo phì với nhau')
else:
    print(f'Trị số p = {p:} >= {alpha:.4f} cho nên không có sự khác biệt trung bình độ tuổi giữa các mức độ béo phì với nhau')
```

→ Kết quả:

```
ANOVA Test: F-statistic = 77.95, p-value = 0.000
Trị số p = 3.59257951669993e-88 < 0.0500 cho nên bác bỏ H0 ==> Có sự khác biệt giữa trung bình độ tuổi và các mức độ béo phì với nhau
```

Hình 18: Kết quả kiểm định của biến age

F-statistic	77.95
p-value	3.59×10^{-88}

Bảng 2. Kết quả kiểm định của biến age

- Vì $p = 3.59 \times 10^{-88} < 0.05$, giả thuyết H_0 bị bác bỏ. Điều này nghĩa là trung bình độ tuổi giữa các nhóm phân loại mức độ béo phì có sự khác biệt. Ví dụ:
  - Người trẻ có thể có xu hướng thừa cân nhẹ (Overweight Level I & II).
  - Người lớn tuổi hơn có thể thuộc nhóm béo phì cấp độ cao (Obesity Type II & III).
- Độ tuổi là một yếu tố có ý nghĩa thống kế đến mức độ béo phì.

Kiểm định 2: *Tần suất tiêu thụ rau củ trong các bữa ăn (vegetables) không ảnh hưởng đến Obesity*

- H_0 : Tần suất xuất hiện rau củ trong bữa ăn là như nhau giữa các mức độ béo phì.
- H_1 : Có sự khác biệt về tần suất xuất hiện rau củ trong bữa ăn giữa các mức độ béo phì với nhau.

→ Kết quả:

```
ANOVA Test: F-statistic = 91.02, p-value = 0.000
Trị số p = 0.0000 < 0.0500 cho nên bác bỏ H0 ==> Có sự khác biệt tần suất xuất hiện rau củ trong bữa ăn giữa các mức độ béo phì với nhau
```

Hình 19. Kết quả kiểm định của biến vegetables

F-statistic	91.02
p-value	0.000

Bảng 3. Kết quả kiểm định của biến vegetables

- Vì $p < 0.05$, giả thuyết H_0 bị bác bỏ. Điều này nghĩa là có sự khác biệt ý nghĩa về tần suất xuất hiện rau củ trong bữa ăn giữa các mức độ béo phì.
- Tần suất ăn rau củ có liên quan đến mức độ béo phì. Những người tiêu thụ nhiều rau củ có thể nằm trong nhóm cân nặng bình thường hoặc thừa cân nhẹ, trong khi những người ít tiêu thụ rau củ thường có nguy cơ béo phì cao hơn.
- Điều này gợi ý rằng chế độ ăn giàu rau củ có thể đóng vai trò quan trọng trong việc kiểm soát cân nặng và phòng ngừa béo phì.

Kiểm định 3: *Chiều cao (height) không ảnh hưởng đến Obesity*

- H_0 : Trung bình chiều cao là như nhau giữa các mức độ béo phì.

- H_1 : Có sự khác biệt về trung bình chiều cao giữa các mức độ béo phì với nhau.

→ Kết quả:

```
ANOVA Test: F-statistic = 38.43, p-value = 0.000
Trị số p = 1.6858535844061656e-44 < 0.0500 cho nên bác bỏ H0 ==> Có sự khác biệt trung bình chiều cao giữa các mức độ béo phì với nhau
```

Hình 20. Kết quả kiểm định của biến height

F-statistic	38.43
p-value	1.69×10^{-44}

Bảng 4. Kết quả kiểm định của biến height

- Vì $p = 1.69 \times 10^{-44} < 0.05$, giả thuyết H_0 bị bác bỏ. Điều này nghĩa là có sự khác biệt ý nghĩa về chiều cao trung bình giữa các mức độ béo phì.
- Sự khác biệt về chiều cao có thể phản ánh tác động của yếu tố cơ địa hoặc gen di truyền đối với nguy cơ béo phì. Ví dụ, người có chiều cao thấp hơn có thể có nguy cơ béo phì cao hơn do tỷ lệ cơ thể nhỏ hơn và ít cơ hội đốt cháy năng lượng hơn.
- Điều này cũng cho thấy cần phân tích thêm mối quan hệ giữa chiều cao, cân nặng và mức độ hoạt động thể chất.

Kiểm định 4: *Số lượng bữa ăn chính (main meals) không ảnh hưởng đến Obesity*

- H_0 : Trung bình số bữa ăn chính là như nhau giữa các mức độ béo phì.
- H_1 : Có sự khác biệt về trung bình số bữa ăn chính giữa các mức độ béo phì với nhau.

→ Kết quả:

```
ANOVA Test: F-statistic = 25.30, p-value = 0.000
Trị số p = 3.8550490118623074e-29 < 0.0500 cho nên bác bỏ H0 ==> Có sự khác biệt trung bình số bữa ăn giữa các mức độ béo phì với nhau
```

Hình 21. Kết quả kiểm định của biến main meals

F-statistic	25.30
p-value	3.85×10^{-29}

Bảng 5. Kết quả kiểm định của biến main meals

- Vì $p = 3.85 \times 10^{-29} < 0.05$, giả thuyết H_0 bị bác bỏ. Điều này nghĩa là có sự khác biệt ý nghĩa về số lượng bữa ăn chính trung bình giữa các mức độ béo phì.
- Kết quả này chỉ ra rằng số lượng bữa ăn chính có liên quan đến béo phì. Những người ăn nhiều bữa chính hơn có thể có xu hướng kiểm soát lượng ăn uống tốt hơn so với người ăn ít bữa nhưng ăn nhiều hơn trong mỗi bữa.
- Cần cân nhắc thêm các yếu tố khác như lượng calo tiêu thụ mỗi bữa và mức độ vận động sau ăn để hiểu rõ hơn vai trò của thói quen ăn uống.

Kiểm định 5: *Lượng nước uống mỗi ngày (water) không ảnh hưởng đến Obesity*

- H_0: Trung bình lượng nước uống mỗi ngày là như nhau giữa các mức độ béo phì.
- H_1: Có sự khác biệt về lượng nước uống mỗi ngày giữa các mức độ béo phì với nhau.

→ Kết quả:

```
ANOVA Test: F-statistic = 17.66, p-value = 0.000
Trị số p = 4.73705000035435097e-20 < 0.0500 cho nên bác bỏ H0 ==> Có sự khác biệt trung bình lượng nước uống mỗi ngày giữa các mức độ béo phì với nhau
```

Hình 22. Kết quả kiểm định của biến water

F-statistic	17.66
p-value	4.74×10^{-20}

Bảng 6. Kết quả kiểm định của biến water

- Vì $p = 4.74 \times 10^{-20} < 0.05$, giả thuyết H_0 bị bác bỏ. Điều này nghĩa là có sự khác biệt ý nghĩa về lượng nước uống mỗi ngày giữa các mức độ béo phì.
- Uống nước đủ có thể giúp giảm nguy cơ béo phì bằng cách tăng cường trao đổi chất, giảm lượng calo tiêu thụ và hỗ trợ cảm giác no.
- Nhóm béo phì nặng có thể uống nước ít hơn hoặc thay thế nước bằng các loại đồ uống có đường, góp phần làm tăng lượng calo hấp thụ.

Kiểm định 6: *Tần suất hoạt động thể thao (physical activity) không ảnh hưởng đến Obesity*

- H_0: Tần suất hoạt động thể thao là như nhau giữa các mức độ béo phì.
- H_1: Có sự khác biệt về tần suất hoạt động thể thao giữa các mức độ béo phì với nhau.

→ Kết quả:

```
ANOVA Test: F-statistic = 17.20, p-value = 0.000
Trị số p = 1.6807840980074483e-19 < 0.0500 cho nên bác bỏ H0 ==> Có sự khác biệt trung bình chiều cao giữa các mức độ béo phì với nhau
```

Hình 23. Kết quả kiểm định của biến physical activity

F-statistic	17.20
p-value	1.68×10^{-19}

Bảng 7. Kết quả kiểm định của biến physical activity

- Vì $p = 1.68 \times 10^{-19} < 0.05$, giả thuyết H_0 bị bác bỏ. Điều này nghĩa là có sự khác biệt ý nghĩa về tần suất hoạt động thể thao giữa các mức độ béo phì.
- Hoạt động thể thao đóng vai trò quan trọng trong việc kiểm soát cân nặng. Nhóm có mức độ béo phì cao có thể ít tham gia vào hoạt động thể thao hơn do thói quen ít vận động hoặc các vấn đề sức khỏe.
- Khuyến khích tăng cường vận động thường xuyên, đặc biệt là các bài tập vừa phải, có thể giúp cải thiện sức khỏe toàn diện và kiểm soát cân nặng.

Kiểm định 7: *Thời gian sử dụng các thiết bị điện tử (tech devices) mỗi ngày không ảnh hưởng đến Obesity*

- H_0: Trung bình thời gian sử dụng các thiết bị điện tử mỗi ngày là như nhau giữa các mức độ béo phì.
- H_1: Có sự khác biệt về trung bình thời gian sử dụng các thiết bị điện tử mỗi ngày giữa các mức độ béo phì với nhau.

→ Kết quả:

```
ANOVA Test: F-statistic = 7.694943122726869, p-value = 3.375625780587961e-08
Kruskal-Wallis Test: H-statistic = 49.86649850793875, p-value = 4.9998455090557935e-09
Trị số p = 3.375625780587961e-08 < 0.0500 cho nên bác bỏ H0 ==> Có sự khác biệt trung bình chiều cao giữa các mức độ béo phì với nhau
```

Hình 24. Kết quả kiểm định của biến tech_devices

F-statistic	7.69
p-value (ANOVA)	3.38×10^{-8}
p-value (Kruskal-Wallis)	4.99×10^{-9}

Bảng 8. Kết quả kiểm định của biến tech_devices

- Vì $p < 0.05$, giả thuyết H_0 bị bác bỏ. Điều này nghĩa là có sự khác biệt ý nghĩa về thời gian sử dụng thiết bị điện tử mỗi ngày giữa các mức độ béo phì.
- Thời gian sử dụng thiết bị điện tử dài hơn có thể liên quan đến lối sống ít vận động, làm tăng nguy cơ béo phì.
- Tuy nhiên, ảnh hưởng này cũng có thể phức tạp, tùy thuộc vào cách sử dụng thiết bị (ví dụ: làm việc, học tập hay giải trí). Việc giảm thời gian sử dụng thiết bị điện tử và kết hợp hoạt động thể chất có thể giúp giảm nguy cơ béo phì.

3. Kiểm định các biến định danh với biến phụ thuộc

Sau khi hoàn thành kiểm định các biến định lượng, nhóm sẽ tiếp tục kiểm định với các thuộc tính phân loại, bao gồm các biến là ['gender', 'transportation', 'alcohol', 'high_caloric', 'family_history', 'between_meals']. Nhóm sẽ sử dụng kiểm định Chi-bình phương giữa các biến độc lập với biến độc lập và biến mục tiêu.

Kiểm định 8: *Giới tính và BMI độc lập*

- H_0 : μ(BMI)[Male] = μ(BMI)[Female]
- H_1 : μ(BMI)[Male] != μ(BMI)[Female]

Xác nhận số lượng mẫu trong mỗi nhóm (nam và nữ), đảm bảo rằng dữ liệu đủ lớn để thực hiện kiểm định.

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

Phương sai được tính để xem xét liệu hai nhóm có sự khác biệt đáng kể về biến động dữ liệu hay không, để quyết định sử dụng phép kiểm định với giả định phương sai bằng nhau hoặc không.

```
df_gb = df.groupby('male')['BMI'].var()
print(df_gb)
```

weightstats.DescrStatsW tạo đối tượng thống kê mô tả cho từng nhóm.

ztest_ind(usevar='unequal') thực hiện kiểm định z-test với giả định phương sai không bằng nhau (Welch's t-test). Kết quả được so sánh với mức ý nghĩa α=0.05 để bác bỏ hoặc không bác bỏ H0.

```
male_bmi = df[df['male'] == 1]['BMI']
female_bmi = df[df['male'] == 0]['BMI']

male_bmi_array = np.array(male_bmi)
female_bmi_array = np.array(female_bmi)

alpha = 0.05
confidence_level = 1 - alpha

col1 = weightstats.DescrStatsW(male_bmi)
col2 = weightstats.DescrStatsW(female_bmi)
cm_obj = weightstats.CompareMeans(col1, col2)
zstat, pvalue = cm_obj.ztest_ind(usevar='unequal')

if (pvalue < alpha):
    print(f'Trị số p = {pvalue:.4f} < {alpha}',
          'nên bác bỏ H0.\n=> μ[Male] != μ[Female]')
else:
    print(f'Trị số p = {pvalue:.4f} >= {alpha}',
          'nên chấp nhận H0.\n=> μ[Male] == μ[Female]')
```

→ Kết quả:

```
Trị số p = 0.0152 < 0.05 nên bác bỏ H0.
=> μ[Male] != μ[Female]
```

Hình 25. Kết quả kiểm định của biến gender

- Vì $p = 0.0152 < 0.05$, giả thuyết H_0 bị bác bỏ. Do đó, có sự khác biệt trung bình BMI giữa nam và nữ. Kết quả này cho thấy rằng giới tính có ảnh hưởng đến chỉ số BMI, với một số lý do tiềm năng như sau:
- Nam giới thường có khối lượng cơ nhiều hơn và tỷ lệ trao đổi chất cơ bản (BMR) cao hơn nữ giới, dẫn đến khác biệt tự nhiên về BMI.
- Nam giới có xu hướng tiêu thụ nhiều calo hơn, trong khi nữ giới có thể ưu tiên kiểm soát cân nặng hoặc chế độ ăn uống lành mạnh hơn.
- Vì BMI phụ thuộc vào cả cân nặng và chiều cao, sự khác biệt về chiều cao trung bình giữa nam và nữ cũng góp phần vào sự chênh lệch này.

Kiểm định 9: *Phương tiện thường sử dụng để di chuyển và BMI độc lập*

- H_0: μ(BMI)[Public_Transportation] = μ(BMI)[Automobile] = μ(BMI)[Walking] = μ(BMI)[Motorbike] = μ(BMI)[Bike]
- H_1: tồn tại ít nhất một μ khác biệt

→ Kết quả:

```
Trị số p = 9.93929385311548e-09 < 0.0500 cho nên bác bỏ H0 ==> có sự khác biệt giữa các khu vực
```

Hình 26. Kết quả kiểm định của biến transportation

- Vì $p = 9.94 \times 10^{-9} < 0.05$, giả thuyết H_0 bị bác bỏ. Điều này nghĩa là có sự khác biệt về chỉ số trung bình BMI giữa các nhóm phương tiện di chuyển (Public Transportation, Automobile, Walking, Motorbike, và Bike).
- Kết quả cho thấy các nhóm người sử dụng các phương tiện di chuyển khác nhau (như ô tô, xe máy, đi bộ, hoặc xe đạp) có mức độ khác biệt về cân nặng hoặc chiều cao, dẫn đến khác biệt về BMI. Có thể vì một số lý do như sau:
- Những người di chuyển bằng phương tiện như đi bộ hoặc xe đạp có xu hướng hoạt động thể chất nhiều hơn, giúp duy trì cân nặng hợp lý và giảm BMI.
- Người sử dụng ô tô hoặc xe máy thường ít hoạt động thể chất trong quá trình di chuyển, có khả năng dẫn đến BMI cao hơn.

- Lựa chọn phương tiện di chuyển cũng có thể phản ánh điều kiện kinh tế xã hội, ảnh hưởng đến chế độ ăn uống, lối sống và BMI. Ví dụ, người sử dụng ô tô thường có thu nhập cao hơn, có thể tiêu thụ thực phẩm giàu calo hơn.

Kiểm định 10: *Tần suất sử dụng đồ uống có chứa cồn không có ảnh hưởng tới mức độ béo phì*

- H_0: Tần suất sử dụng đồ uống có chứa cồn và mức độ béo phì độc lập.
- H_1: Tần suất sử dụng đồ uống có chứa cồn có tác động đến mức độ béo phì.

Tạo crosstable chứa số lần xuất hiện của từng kết hợp giữa các mức độ béo phì (obesity) và tần suất sử dụng đồ uống có cồn (alcohol). Đây là dữ liệu đầu vào cần thiết cho kiểm định Chi-square.

```
# Tạo bảng tần suất (Contingency Table)
contingency_table_2 = pd.crosstab(df['obesity'], df['alcohol'])
print("Contingency Table:")
print(contingency_table_2)
```

Kết quả bảng:

Contingency Table:

alcohol	Always	Frequently	Sometimes	no
obesity				
1	0	1	154	117
2	1	18	161	107
3	0	16	224	50
4	0	19	143	128
5	0	14	172	165
6	0	2	224	71
7	0	0	323	1

Hình 27. Bảng tần suất của biến alcohol và biến obesity

Tính giá trị thống kê Chi-square (χ^2), p-value, bậc tự do (*df*), và giá trị kỳ vọng. Kết quả của kiểm định cho phép xác định liệu hai biến có mối quan hệ độc lập hay không.

```
## Kiểm định Chi-square
stat, p, ddof, expected = stats.chi2_contingency(contingency_table_2)
```

So sánh giá trị p-value với ngưỡng ý nghĩa (α) để đưa ra kết luận. Xác định xem có đủ bằng chứng để bác bỏ giả thuyết H0 hay không.

```
# critical = stats.chi2.ppf(confidence_level, ddof)
# print(f'probability={confidence_level:.4f}\ncritical={critical:.4f}\nstat={stat:.3f}\n')

if (p < alpha):
   print(f'Trị số p = {p} < {alpha:.4f} nên bác bỏ H0.',
       '\n=> Tần suất sử dụng đồ uống có chứa cồn  có ảnh hưởng đến trạng thái béo phì.')
else:
   print(f'Trị số p = {p} >= {alpha:.4f} nên chấp nhận H0.',
       '\n=> Tần suất sử dụng đồ uống có chứa cồn không ảnh hưởng đến trạng thái béo phì')
```

→ Kết quả:

```
Trị số p = 5.287157877798169e-61 < 0.0500 nên bác bỏ H0.
=> Tần suất sử dụng đồ uống có chứa cồn  có ảnh hưởng đến trạng thái béo phì.
```

Hình 28. Kết quả kiểm định của biến alcohol

- Vì $p = 5.29 \times 10^{-61} < 0.05$, giả thuyết H_0 bị bác bỏ. Tần suất sử dụng đồ uống có cồn có mối liên hệ với mức độ béo phì. Có thể giải thích bởi một số nguyên nhân như sau:
- Nhiều loại đồ uống có cồn (ví dụ: bia, cocktail) chứa lượng calo cao, dẫn đến tích lũy năng lượng thừa.
- Việc sử dụng đồ uống có cồn có thể làm giảm khả năng kiểm soát ăn uống hoặc thúc đẩy việc tiêu thụ thực phẩm giàu calo đi kèm.
- Người thường xuyên uống rượu bia có thể ít tham gia vào các hoạt động thể chất hơn, làm tăng nguy cơ béo phì.

Kiểm định 11: *Sử dụng thức ăn chứa nhiều calo không có ảnh hưởng tới mức độ béo phì*

- H_0: Sử dụng thức ăn chứa nhiều calo và mức độ béo phì độc lập.
- H_1: Sử dụng thức ăn chứa nhiều calo có tác động đến mức độ béo phì.

→ Kết quả:

```
Trị số p = 4.2280167944702657e-131 < 0.0500 nên bác bỏ H0.
=> Sử dụng thức ăn chứa nhiều calo có ảnh hưởng đến trạng thái béo phì.
```

Hình 29. Kết quả kiểm định của biến high_caloric

- Vì $p = 4.23 \times 10^{-131} < 0.05$, giả thuyết H_0 bị bác bỏ. Tần suất sử dụng thức ăn chứa nhiều calo có liên hệ chặt chẽ với mức độ béo phì. Có thể giải thích bởi một số nguyên nhân như sau:
- Thức ăn chứa nhiều calo như thức ăn nhanh, đồ chiên, thực phẩm chế biến sẵn thường dẫn đến việc cung cấp năng lượng vượt quá nhu cầu cơ thể.
- Khi năng lượng tiêu thụ cao hơn năng lượng tiêu hao, cơ thể tích trữ mỡ thừa, gây tăng cân và béo phì.
- Những thực phẩm giàu calo nhưng nghèo dinh dưỡng cũng góp phần làm tăng nguy cơ béo phì vì chúng không cung cấp cảm giác no kéo dài, dễ dẫn đến ăn quá mức.

Kiểm định 12: *Tiền sử gia đình có người mắc béo phì không có ảnh hưởng tới mức độ béo phì*

- H_0 : Gia đình đã có người mắc béo phì và mức độ béo phì độc lập.
- H_1 : Gia đình đã người mắc béo phì có tác động đến mức độ béo phì.

→ **Kết quả:**

```
Trị số p = 4.2280167944702657e-131 < 0.0500 nên bác bỏ H0.
=> Tần suất sử dụng đồ uống có chứa cồn  có ảnh hưởng đến trạng thái béo phì.
```

Hình 30. Kết quả kiểm định của biến family_history

- Vì $p = 4.23 \times 10^{-131} < 0.05$, giả thuyết H_0 bị bác bỏ. Một yếu tố di truyền hoặc môi trường gia đình có ảnh hưởng mạnh mẽ đến béo phì:
- Nhiều nghiên cứu đã chỉ ra rằng gen ảnh hưởng đến tốc độ chuyển hóa, cảm giác no, và xu hướng lưu trữ chất béo.
- Những người sống trong gia đình có thói quen ăn uống không lành mạnh hoặc ít vận động cũng có nguy cơ cao mắc béo phì.
- Dù yếu tố di truyền quan trọng, môi trường gia đình như chế độ ăn uống và hoạt động thể chất cũng đóng vai trò lớn trong việc xác định mức độ béo phì.

Kiểm định 13: *Tần suất ăn vặt (các bữa ăn khác ngoài bữa chính) không có ảnh hưởng tới mức độ béo phì*

- H_0: Tần suất ăn vặt (các bữa ăn khác ngoài bữa chính) và mức độ béo phì độc lập.
- H_1: Tần suất ăn vặt (các bữa ăn khác ngoài bữa chính) có tác động đến mức độ béo phì.

→ Kết quả:

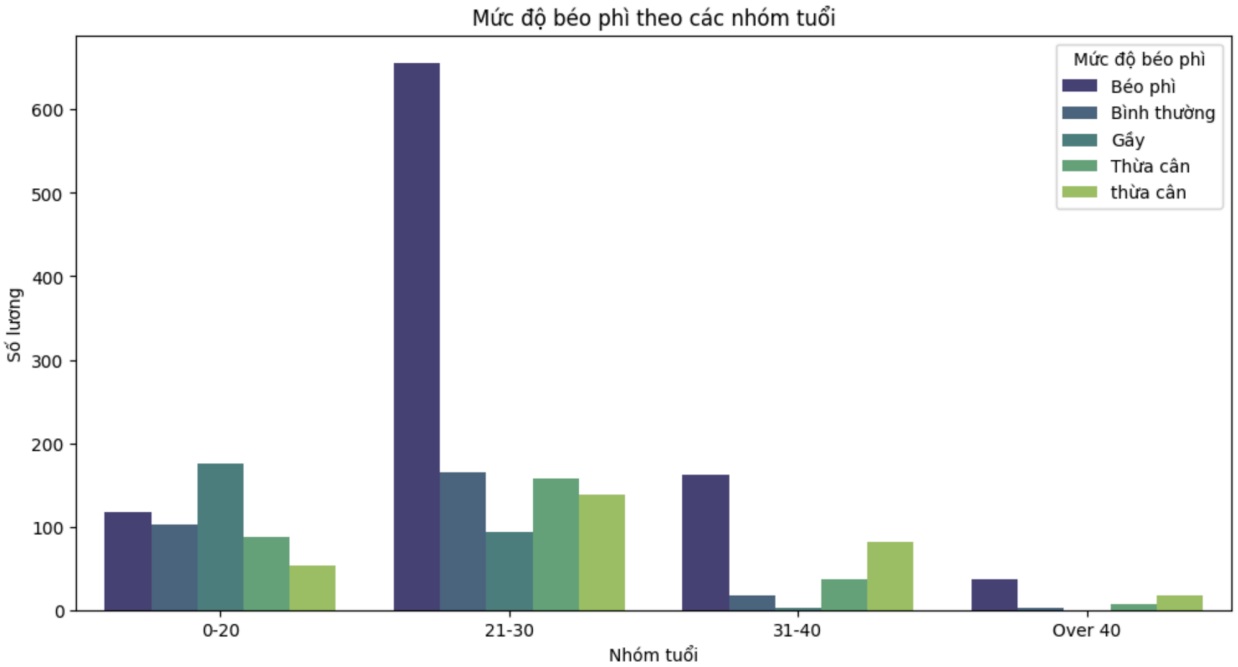
```
Trị số p = 7.383852893286775e-159 < 0.0500 nên bác bỏ H0.
=> Tần suất sử dụng đồ uống có chứa cồn  có ảnh hưởng đến trạng thái béo phì.
```

Hình 31. Kết quả kiểm định của biến bewteen_meals

- Vì $p = 7.38 \times 10^{-159} < 0.05$, giả thuyết H_0 bị bác bỏ. Tần suất ăn vặt có liên hệ mạnh mẽ với mức độ béo phì, điều này cũng khá dễ hiểu vì:
- Đồ ăn vặt thường có lượng calo cao, ít chất xơ, giàu đường và chất béo bão hòa, làm tăng nguy cơ tích lũy năng lượng thừa.
- Thường xuyên ăn vặt có thể thay thế các bữa ăn chính cân bằng, dẫn đến thiếu hụt vi chất dinh dưỡng nhưng dư thừa năng lượng.
- Tần suất ăn vặt cao thường phổ biến ở môi trường có sẵn nhiều đồ ăn tiện lợi và thói quen ăn uống ít kiểm soát.

Chương 5. Trực quan hóa dữ liệu

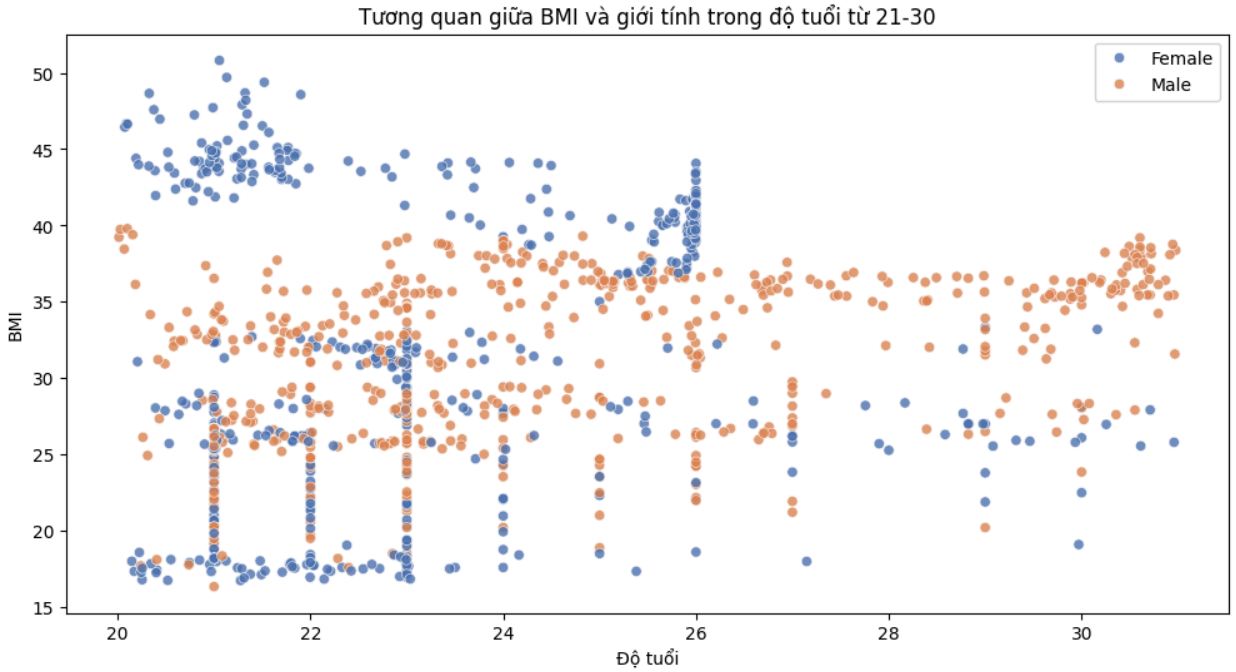
1. Mức độ béo phì theo các nhóm tuổi



Hình 32. Biểu đồ mức độ béo phì theo các nhóm tuổi

Nhìn chung, số lượng người béo phì đạt đỉnh ở nhóm tuổi 21-30 và giảm dần khi tuổi tăng. Đây là nhóm có số lượng người béo phì cao nhất so với các nhóm tuổi khác. Điều này có thể lý giải do dữ liệu thu thập được trong nhóm này nhiều hơn các nhóm khác và thói quen sống hoặc chế độ ăn uống ít lành mạnh của nhóm tuổi này. Ở độ tuổi 0 đến 20, tỷ lệ gầy, bình thường, thừa cân, béo phì khá cân bằng. Điều này có thể lý giải rằng nhóm này thường được kiểm soát chế độ dinh dưỡng bởi gia đình hoặc nhà trường. Với 2 nhóm tuổi 31 - 40 và trên 40, tỷ lệ người béo phì vẫn chiếm tỷ lệ cao nhất.

2. Tương quan giữa BMI và giới tính trong độ tuổi từ 21-30



Hình 35. Biểu đồ tương quan giữa BMI và giới tính trong độ tuổi từ 21-30

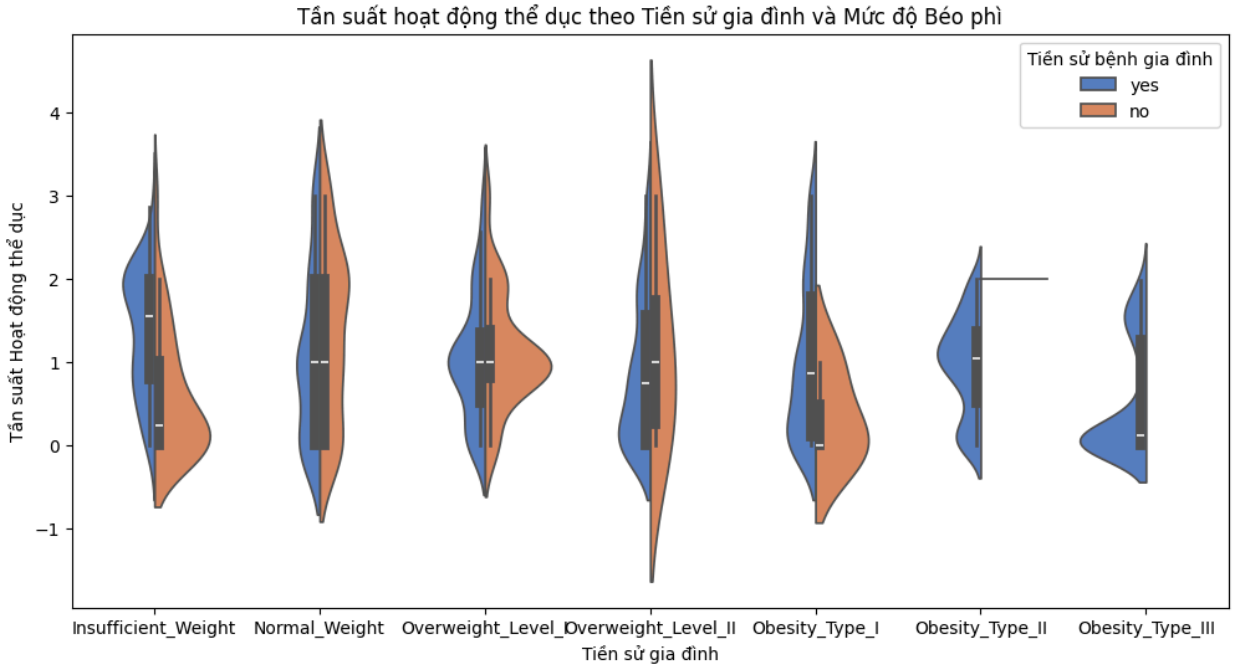
Biểu đồ scatter plot thể hiện sự phân bổ BMI (chỉ số khối cơ thể) theo độ tuổi từ 21 đến 30, phân biệt giữa hai giới tính, giúp ta có cái nhìn rõ hơn về xu hướng béo phì ở mỗi nhóm. Qua biểu đồ, ta thấy rằng BMI ở nữ có sự trải rộng trên toàn dải giá trị, với một số điểm rải rác từ mức BMI thấp đến rất cao. Điều này cho thấy rằng nữ giới trong độ tuổi này có sự đa dạng rõ rệt về trạng thái sức khỏe, từ những người có BMI trong mức bình thường đến những người có BMI thuộc dạng thừa cân hoặc béo phì. Trong khi đó, ở nam giới, dữ liệu BMI chủ yếu phân bổ trong khoảng giá trị từ 25 đến 40, cho thấy rằng nam giới có xu hướng có BMI ổn định hơn, phần lớn nằm ở mức bình thường hoặc thừa cân nhẹ.

Đặc biệt, biểu đồ chỉ ra rằng nữ giới trong độ tuổi từ 20 đến 26 tuổi có xu hướng đạt BMI cao hơn so với nam giới trong cùng độ tuổi. Sự phân bổ này có thể phản ánh những yếu tố sinh lý, hormon, và thói quen sinh hoạt ảnh hưởng mạnh đến nữ giới. Các nghiên cứu trước đây đã chỉ ra rằng phụ nữ thường có tỷ lệ mỡ cơ thể cao hơn nam giới, đồng thời sự thay đổi hormon trong độ tuổi này cũng có thể góp phần làm tăng nguy cơ thừa cân hoặc béo phì.

Ngược lại, nam giới trong độ tuổi này lại có xu hướng có mức BMI thấp hoặc trung bình phổ biến hơn, và ít có dấu hiệu của béo phì so với nữ giới. Điều này có thể do nam giới thường có tỷ lệ cơ bắp cao hơn và dễ dàng duy trì một lối sống năng động hơn. Tuy nhiên, cũng cần lưu ý rằng các yếu tố như thói quen ăn uống, chế độ luyện tập thể dục thể thao, cũng như thói quen sống có thể đóng vai trò quan trọng trong việc hình thành sự khác biệt này.

Nhìn chung, từ biểu đồ, ta có thể nhận thấy rằng nữ giới trong độ tuổi từ 21 đến 30 có khả năng bị béo phì (BMI cao) cao hơn so với nam giới. Các yếu tố có thể giải thích cho sự khác biệt này có thể bao gồm chế độ ăn uống, thói quen vận động, lối sống hàng ngày và sự ảnh hưởng của các yếu tố hormon. Vì vậy, việc tăng cường nhận thức về lối sống lành mạnh, cải thiện chế độ ăn uống và khuyến khích hoạt động thể chất có thể là những biện pháp hữu hiệu để giảm thiểu tỷ lệ béo phì trong nhóm nữ giới ở độ tuổi này.

3. Phân phối tần suất hoạt động thể dục thể thao theo tiền sử gia đình và mức độ béo phì



Hình 33. Biểu đồ tần suất hoạt động TDTT theo tiền sử gia đình và mức độ béo phì

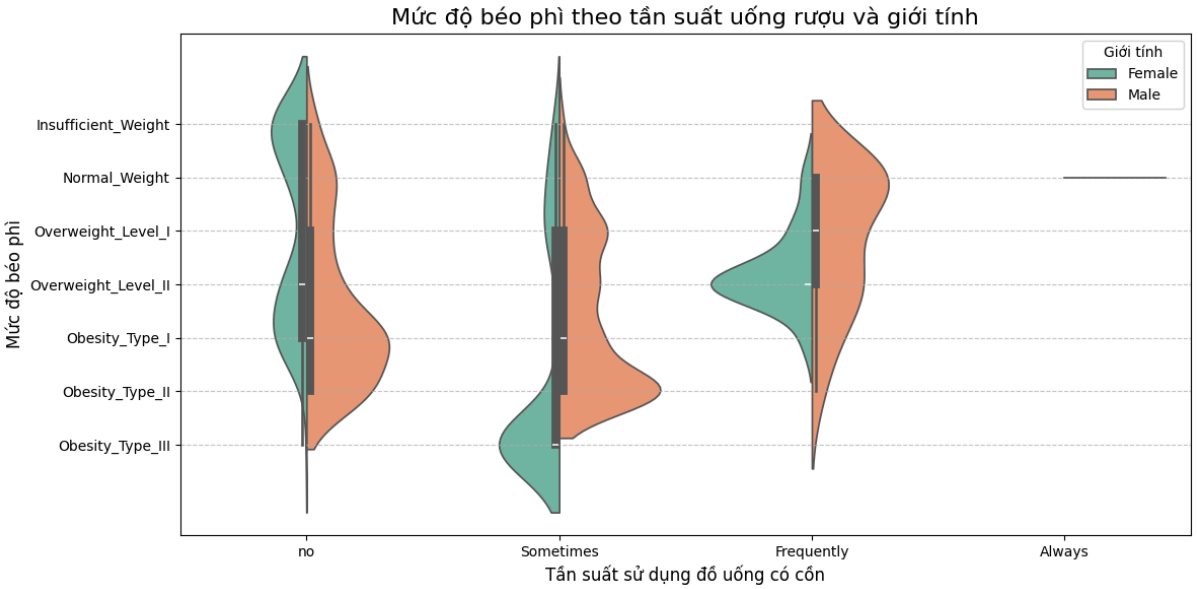
Từ biểu đồ violin, nhóm cân nặng bình thường (*Normal_Weight*) cho thấy tần suất hoạt động thể dục đều hơn so với các nhóm khác và phân bố trải rộng, phản ánh rằng những người trong nhóm này thường duy trì mức độ vận động tốt, bất kể có hay không có tiền sử bệnh gia đình.

Trong khi đó, ở nhóm thừa cân trở về sau, tần suất hoạt động thể dục giảm dần, đặc biệt rõ ở nhóm có tiền sử bệnh gia đình (màu xanh). Ở các mức béo phì nặng (*Obesity_Type II* và *Obesity_Type III*), tần suất hoạt động thể dục rất thấp hơn khá nhiều, cho thấy sự giảm sút đáng kể trong hoạt động thể chất của những người thuộc các nhóm này.

Nhóm không có tiền sử bệnh gia đình (màu cam) trong tình trạng thừa cân II có xu hướng hoạt động thể dục cao hơn so với nhóm có tiền sử bệnh gia đình. Có thể suy đoán những người không có tiền sử béo phì trong gia đình sẽ quan tâm tới tình trạng kiểm soát cân nặng khi bắt đầu có dấu hiệu thừa cân. Những người không có tiền sử béo phì trong gia đình cũng sẽ không xuất hiện trong nhóm béo phì nặng (*Obesity_Type II* và *Obesity_Type III*).

Từ đây, ta có thể nhận ra rằng những người không có người thân trong gia đình bị béo phì, tần suất hoạt động thể dục thể thao giúp họ giảm nguy cơ mắc béo phì. Ngược lại, những người thừa cân, béo phì và có người thân trong gia đình cũng bị thừa cân, béo phì sẽ có xu hướng không thường xuyên hoạt động thể dục thể thao (mà tần suất hoạt động TDTT lại có ảnh hưởng đến mức độ béo phì). Vì vậy, có thể kết luận rằng: nếu một vài thành viên trong gia đình bị thừa cân, béo phì do không tham gia thể thao sẽ khiến các thành viên không tích cực tham gia thể thao, từ đó khiến họ dễ bị thừa cân, béo phì.

4. Mức độ béo phì theo tần suất uống rượu và giới tính



Hình 34. Biểu đồ mức độ béo phì theo tần suất uống rượu và giới tính

Biểu đồ violin cho thấy mối quan hệ giữa mức độ béo phì và tần suất sử dụng đồ uống có cồn theo giới tính, mang đến cái nhìn sâu sắc về thói quen tiêu thụ đồ uống có cồn và sự phân bố trạng thái béo phì ở nam và nữ. Nhìn chung, biểu đồ cho thấy nam giới có xu hướng sử dụng đồ uống có cồn nhiều hơn nữ giới, đặc biệt là ở các nhóm thừa cân và béo phì.

Đối với nhóm không sử dụng đồ uống có cồn ("no"), tỷ lệ nữ béo phì thấp hơn đáng kể so với nam giới. Phần lớn nữ giới tập trung ở nhóm "Insufficient_Weight" (thiếu cân) hoặc "Normal_Weight" (cân nặng bình thường), trong khi nam giới chủ yếu nằm ở các mức độ béo phì, đặc biệt là "Obesity_Type_I" và "Obesity_Type_II". Điều này chỉ ra rằng nam giới không sử dụng đồ uống có cồn có xu hướng có mức độ béo phì cao hơn nữ giới.

Nhóm thỉnh thoảng sử dụng đồ uống có cồn ("Sometimes") có sự phân bổ tần suất khá đồng đều giữa nam và nữ ở các mức cân nặng từ bình thường đến thừa cân. Tuy nhiên, tỷ lệ béo phì vẫn cao hơn ở nam giới, với sự tập trung ở các mức "Overweight_Level_I" và "Overweight_Level_II". Điều này cho thấy việc sử dụng đồ uống có cồn, mặc dù không phải là thói quen thường xuyên, vẫn có thể ảnh hưởng đến trạng thái cân nặng và mức độ béo phì của người sử dụng, đặc biệt là ở nam giới.

Với nhóm sử dụng đồ uống có cồn thường xuyên ("Frequently"), nam giới có xu hướng trải rộng ở tất cả các mức độ béo phì, từ thiếu cân đến béo phì nghiêm trọng. Trong khi đó, nữ giới chủ yếu tập trung ở nhóm "Normal_Weight" hoặc "Overweight_Level_I" (thừa cân mức độ I). Đây là dấu hiệu cho thấy nam giới có sự ảnh hưởng mạnh mẽ từ thói quen uống rượu đối với mức độ béo phì của họ, đặc biệt là khi sử dụng đồ uống có cồn thường xuyên.

Tóm lại, đối với nữ giới, việc sử dụng đồ uống có cồn sẽ khiến nữ giới rơi vào tình trạng thừa cân và béo phì. Trong khi đó, nam giới không hoặc càng ít sử dụng đồ uống có cồn lại có nguy cơ thừa cân béo phì cao hơn những người sử dụng thường xuyên.

TÀI LIỆU THAM KHẢO

- Github: https://github.com/hnamtraan/DataVisualizationUEH/tree/main
- https://www.kaggle.com/code/tkunzler/obesity-levels-acc-98-eda