#import libary

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

Dữ liệu được cung cấp trong file income\_spending.csv, chứa thông tin khảo sát về thu nhập, thói quen chi tiêu và tiết kiệm của một nhóm người dân. Bộ dữ liệu bao gồm các cột sau:

* age: Tuổi (năm)
* gender: Giới tính (nam = 'M', nữ = 'F')
* income: Thu nhập hàng năm (USD)
* expenses: Chi tiêu hàng tháng (USD)
* savings: Số tiền tiết kiệm hàng tháng (USD)
* education: Trình độ học vấn (High School, Bachelor, Master, PhD)
* marital\_status: Tình trạng hôn nhân (Single, Married, Divorced, Widowed)
* loan\_status: Có đang vay nợ không? (1 = Có, 0 = Không)
* investment: Có đang đầu tư không? (1 = Có, 0 = Không)

### Câu 1: (7 điểm, CLO 1)

#### a. (2 điểm) Tìm hiểu một số thông tin về bộ dữ liệu:

* Hiển thị 5 dòng đầu tiên của bộ dữ liệu. Cho biết dữ liệu có kích thước bao nhiêu dòng và bao nhiêu cột?
* Dữ liệu có giá trị trống không?

#Hiển thị 5 dòng đầu tiên của bộ dữ liệu //("T:/22684961\_PhanNhatQuang/income\_spending.csv")

df = pd.read\_csv("income\_spending.csv")

print(df.head(5))

#Cho biết dữ liệu có kích thước bao nhiêu dòng và bao nhiêu cột? df.shape

print(f'Số dòng: {df.shape[0]}')

print(f'Số cột: {df.shape[1]}')

#Dữ liệu có giá trị trống không? df.isnull().sum()

print("\nSố lượng giá trị thiếu:")

print( df.isnull().sum() )

#### b. (2 điểm) Dựa trên dữ liệu, hãy tính toán hoặc so sánh một số thông tin sau:

* Tính giá trị trung bình, trung vị và độ lệch chuẩn của thu nhập (income) theo từng nhóm giới tính.
* So sánh mức chi tiêu trung bình chi tiêu giữa các nhóm có tình trạng hôn nhân khác nhau (Single, Married, Divorced, Widowed).

mean\_marital\_status= df.groupby("marital\_status")["income"].mean()

mean\_marital\_status

**Tính giá trị trung bình, trung vị và độ lệch chuẩn của thu nhập (income) theo từng nhóm giới tính.**

**So sánh trung bình thu nhập giữa các nhóm có vay nợ (loan\_status = 1) và không vay nợ (loan\_status = 0).**

#Tính giá trị trung bình, trung vị và độ lệch chuẩn của thu nhập (income) theo từng nhóm giới tính.

status\_price\_area = df.groupby("gender")["income"].agg(["mean", "median","std"])

mean= df.groupby("gender")["income"].mean()

median=df.groupby("gender")["income"].median()

std= df.groupby("gender")["income"].std()

print( status\_price\_area )

#So sánh trung bình thu nhập giữa các nhóm có vay nợ (loan\_status = 1) và không vay nợ (loan\_status = 0).

# vẽ biểu đồ tròn: % từng loại bất động sản theo cách thức ra bán và đưa ra ít nhất 2 nhận xét (1.5d)

import pandas as pd

# Tiền xử lý nếu cần

df = df.dropna()

# Tính trung bình, trung vị và độ lệch chuẩn theo giới tính

print("Thống kê theo giới tính:")

thong\_ke\_gt = df.groupby('gender')['income'].agg(['mean', 'median', 'std'])

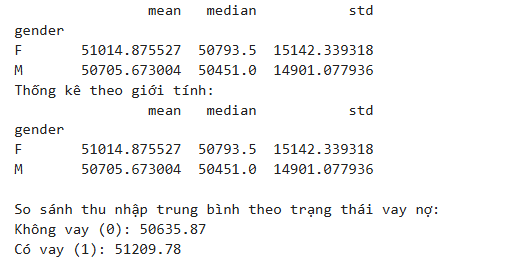
print(thong\_ke\_gt)

print("\nSo sánh thu nhập trung bình theo trạng thái vay nợ:")

thu\_nhap\_vay = df.groupby('loan\_status')['income'].mean()

print("Không vay (0): {:.2f}".format(thu\_nhap\_vay[0]))

print("Có vay (1): {:.2f}".format(thu\_nhap\_vay[1]))



#### c. (2 điểm) Phân tích nhóm đối tượng trung niên với các đặc điểm sau:

* Tuổi từ 30 đến 50
* Thu nhập trên mức trung vị của toàn bộ dữ liệu
* Chi tiêu ít hơn 40% thu nhập hàng tháng

Yêu cẩu:

* Tính số lượng người trong nhóm này và so sánh tỷ lệ của họ so với toàn bộ tập dữ liệu.
* Tính giá trị trung bình của thu nhập, chi tiêu và tiết kiệm của nhóm này

# Trung vị thu nhập toàn bộ

thu\_nhap\_trung\_vi = df['income'].median()

nhom\_trung\_nien = df[

(df['age'] >= 30) &

(df['age'] <= 50) &

(df['income'] > thu\_nhap\_trung\_vi) &

(df['expenses'] < 0.4 \* df['income'])

]

# Tính số lượng và tỷ lệ

so\_luong\_nhom = len(nhom\_trung\_nien)

ty\_le\_nhom = so\_luong\_nhom / len(df) \* 100

print(f"Số lượng người trong nhóm trung niên: {so\_luong\_nhom}")

print(f"Tỷ lệ so với toàn bộ dữ liệu: {ty\_le\_nhom:.2f}%")

# Thêm cột "saving" (tiết kiệm)

# Tạo cột mới "saving" = thu nhập - chi tiêu

nhom\_trung\_nien['saving'] = nhom\_trung\_nien['income'] - nhom\_trung\_nien['expenses']

# Tính trung bình thu nhập, chi tiêu, tiết kiệm

thu\_nhap\_tb = nhom\_trung\_nien['income'].mean()

chi\_tieu\_tb = nhom\_trung\_nien['expenses'].mean()

tiet\_kiem\_tb = nhom\_trung\_nien['saving'].mean()

print(f"\nGiá trị trung bình của nhóm này:")

print(f"Thu nhập trung bình: {thu\_nhap\_tb:.2f}")

print(f"Chi tiêu trung bình: {chi\_tieu\_tb:.2f}")

print(f"Tiết kiệm trung bình: {tiet\_kiem\_tb:.2f}")



#### d. (2 điểm) Vẽ biểu đồ:

* Vẽ biểu đồ tán xạ (scatter plot) với trục X cho thu nhập (income) và trục Y cho chi tiêu (expenses). Tô màu các dấu điểm theo nhóm giới tính.
* Vẽ biểu đồ hộp (box plot) cho thu nhập (income) theo trình độ học vấn (education).

plt.figure(figsize=(8,6))

# plt.scatter(x, y, color, alpha, edgecolor) tạo biểu đồ scatter:

# df["income"]: Trục X (biến độc lập) → Thu nhập.

# df["expenses"]: Trục Y (biến phụ thuộc) → Chi tiêu.

# color="blue": Điểm dữ liệu có màu xanh dương.

# alpha=0.5: Độ trong suốt của điểm (giúp quan sát dữ liệu chồng lấn).

# edgecolor='black': Viền màu đen xung quanh mỗi điểm.

plt.scatter(df["income"],df['expenses'], color="blue", alpha=0.5, edgecolor='black')

# sns.scatterplot(data=df, x="income", y="expenses", hue="gender", alpha=0.5) -- vẽ tô màu theo giới tính với seaborn

# colors = df["gender"].map({"Male": "blue", "Female": "red"}) --vẽ tô màu theo giới tính với matplotlib

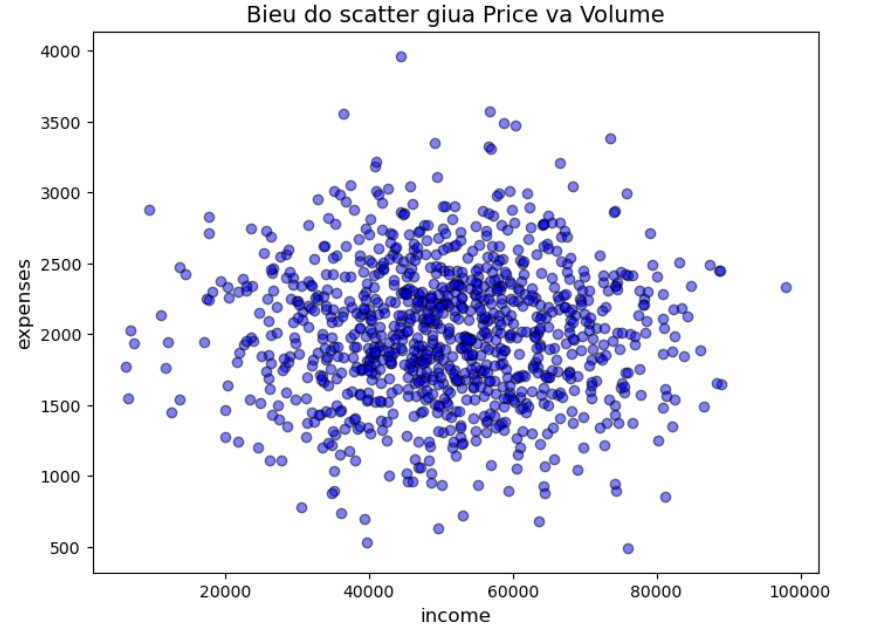
# plt.scatter(df["income"], df["expenses"], c=colors, alpha=0.5)

plt.title("Bieu do scatter giua Price va Volume", fontsize=14)

plt.xlabel(("income"),fontsize=12)

plt.ylabel(("expenses"),fontsize=12)

plt.show()



plt.figure(figsize=(12,6))#Tạo một biểu đồ với kích thước rộng 12, cao 6 để dễ nhìn.

# x="education" → Trục hoành (x) hiển thị trình độ học vấn (education).

# y="income" → Trục tung (y) hiển thị thu nhập (income).

# data=df → Sử dụng dữ liệu từ df.

# Seaborn's boxplot() sẽ tự động vẽ một biểu đồ hộp cho mỗi nhóm trong education.

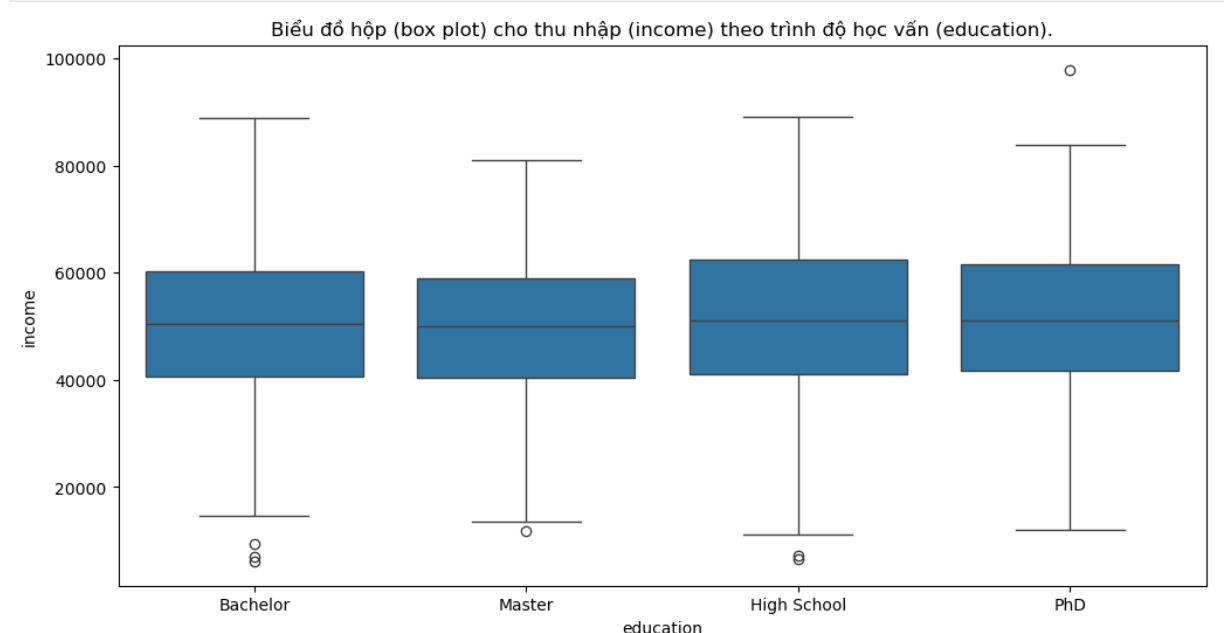
sns.boxplot(x="education",y="income",data=df)

plt.title("Biểu đồ hộp (box plot) cho thu nhập (income) theo trình độ học vấn (education).")

plt.xlabel("education")

plt.ylabel("income")

plt.show()



### **Câu 2: (2 điểm, CLO 2)**

**Sử dụng hàm df.sample() để lấy mẫu ngẫu nhiên 100 dòng từ bộ dữ liệu. Tính giá trị trung bình thu nhập (income) của mẫu này.**

* **Lấy mẫu ngẫu nhiên n = 50 từ bộ dữ liệu và tính trung bình thu nhập trong mẫu. Lặp lại quá trình trên 100 lần và vẽ histogram phân bố của các giá trị trung bình mẫu.**
* **Lặp lại với n = 100, 200 và so sánh kết quả giữa các kích thước mẫu. Nhận xét về sự thay đổi của trung bình mẫu khi kích thước mẫu tăng.**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

# Giả sử df là DataFrame chứa dữ liệu

# Lấy mẫu ngẫu nhiên 100 dòng và tính trung bình thu nhập

sample\_100 = df.sample(n=100, random\_state=42)

mean\_income\_100 = sample\_100['income'].mean()

print(f"Trung bình thu nhập của mẫu 100 dòng: {mean\_income\_100:.2f}")

# Hàm thực hiện lấy mẫu và tính trung bình

def sample\_means(n, num\_samples=100):

means = [df.sample(n=n)['income'].mean() for \_ in range(num\_samples)]

return means

# Lặp lại quá trình với n = 50, 100, 200

sample\_sizes = [50, 100, 200]

plt.figure(figsize=(12,6))

for i, n in enumerate(sample\_sizes):

means = sample\_means(n)

plt.subplot(1, 3, i+1)

plt.hist(means, bins=15, alpha=0.7, color='b', edgecolor='black')

plt.title(f'Kích thước mẫu: {n}')

plt.xlabel('Trung bình thu nhập')

plt.ylabel('Tần suất')

plt.tight\_layout()

plt.show()