BÀI TẬP THỰC HÀNH MÔN NHẬP MÔN PHÂN TÍCH DỮ LIỆU VÀ HỌC SÂU

- Bài tập được thiết kế theo từng lab, mỗi lab là 3 tiết có sự hướng dẫn của GV.
- Cuối mỗi buổi thực hành, sinh viên nộp lại phần bài tập mình đã thực hiện cho GV hướng dẫn.
- Những câu hỏi mở rộng/khó giúp sinh viên trau dồi thêm kiến thức của môn học. Sinh viên phải có trách nhiệm nghiên cứu, tìm câu trả lời nếu chưa thực hiện xong trong giờ thực hành.

LAB 1:

BÀI THỰC HÀNH THAO TÁC DỮ LIỆU

Nội dung: Thao tác dữ liệu điểm thi đại học của học sinh được cho bởi bảng bên dưới

Mục tiêu: Sinh viên đạt được kiến thức sau

1. Tìm hiểu nghiệp vụ dữ liệu
2. Nhập liệu bằng công cụ từ file excel
3. Xác định dữ liệu định tính và định lượng
4. Hiệu chỉnh các thang đo phù hợp và kiểu giá trị dữ liệu cho từng biến số
5. Hiệu chỉnh dữ liệu và xử lý dữ liệu thiếu
6. Chuyển đổi (transformation) dữ liệu theo khoảng cho trước
7. Tạo biến số phụ thuộc theo biến độc lập
8. Tạo biến định tính phân loại

Dữ liệu **dulieuxettuyendaihoc.csv** được mô tả như sau

Dữ liệu lưu trữ điểm trung bình môn, khu vực, khối thi và điểm thi đại học của 100 học sinh.

- T1, L1, H1, S1, V1, X1, D1, N1 lần lượt là điểm trung bình các môn Toán, Lý ,Hóa, Sinh, Văn, Sử, Địa, Ngoại ngữ năm lớp 10
- T2, L2, H2, S2, V2, X2, D2, N2 lần lượt là điểm trung bình các môn Toán, Lý ,Hóa, Sinh, Văn, Sử, Địa, Ngoại ngữ năm lớp 11
- T6, L6, H6, S6, V6, X6, D6, N6 lần lượt là điểm trung bình các môn Toán, Lý ,Hóa, Sinh, Văn, Sử, Địa, Ngoại ngữ năm lớp 12
- GT: Giới tính
- DT: Dân tộc
- KV, KT lần lượt là khu vực thi và khối thi
- DH1, DH2, DH3 lần lượt là điểm thi đại học môn 1, môn 2, môn 3

Sử dụng Pandas để thực hiện các yêu cầu sau đây

1. Xác định và phân loại dữ liệu định tính và định lượng
2. Định nghĩa các thang đo phù hợp cho từng biến số
3. Sử dụng Python để tải dữ liệu lên chương trình và in ra màn hình 10 dòng đầu tiên và 10 dòng cuối cùng
4. Thống kê dữ liệu thiếu cho cột dân tộc và hiệu chỉnh dữ liệu thiếu như sau: Mặc định thiếu thì điền giá trị 0.

Hướng dẫn

1. Lập bảng tần số, tần suất để khảo sát dữ liệu thiếu, bao nhiêu dữ liệu riêng biệt (pandas unique)
2. Thực hiện thay thế dữ liệu thiếu bằng phương pháp điền dữ liệu 0

5. Thống kê dữ liệu thiếu cho biến T1 và hiệu chỉnh dữ liệu, lưu ý việc thay thế dữ liệu thiếu sử dụng phương pháp Mean.

Hướng dẫn

1. Lập bảng tần số, tần suất để khảo sát dữ liệu thiếu

2. Thực hiện thay thế dữ liệu thiếu bằng phương pháp Mean

6. Hãy thực hiện xử lý lần lượt tất cả dữ liệu thiếu cho các biến về điểm số còn lại.
7. Tạo các biến TBM1, TBM2, TBM3 tương ứng với trung bình môn của các năm lớp 10, 11 và 12.
    - Công thức tính: TBM = (T*2 + L + H + S + V*2 + X + D + N) / 10
8. Tạo các biến xếp loại XL1, XL2 và XL3 dựa trên TBM1,TBM2 và TBM3 cho từng năm lớp 10, 11, 12 như sau:
    - Nhỏ hơn 5.0 xếp loại: yếu (kí hiệu là Y)
    - Từ 5.0 đến dưới 6.5: trung bình (kí hiệu là TB)
    - Từ 6.5 đến dưới 8.0: khá (kí hiệu là K)
    - Từ 8.0 đến dưới 9.0: giỏi (kí hiệu là G)
    - Từ 9.0 trở lên: xuất sắc (kí hiệu là XS)
9. Tạo các biến US_TBM1, US_TBM2 và US_TBM3 để chuyển điểm trung bình các năm lớp 10, 11 và 12 từ thang điểm 10 của Việt Nam sang thang điểm 4 của Mỹ. Sử dụng phương pháp Min-Max Normalization
10. Tạo biến kết quả xét tuyển (kí hiệu là KQXT) nhằm xác định sinh viên đậu (giá trị "1") và rớt (giá trị "0") vào các khối dựa trên điểm DH1, DH2 và DH3 như sau
    - Với khối A, A1 nếu [(DH1*2 + DH2 + DH3)/4] lớn hơn hoặc bằng 5.0 thì đậu, ngược lại là rớt
    - Với khối B nếu [(DH1 + DH2*2 + DH3)/4] lớn hơn hoặc bằng 5.0 thì đậu, ngược lại là rớt
    - Với khối khác nếu [(DH1+ DH2 + DH3)/3] lớn hơn hoặc bằng 5.0 thì đậu, ngược lại là rớt
11. Lưu trữ dữ liệu xuống ổ đĩa thành file **processed_dulieuxettuyendaihoc.csv**

LAB 2:

BÀI THỰC HÀNH CHUẨN BỊ DỮ LIỆU

Nội dung: Chuẩn bị dữ liệu – Data Preparation

Tham khảo: seaborn: statistical data visualization — seaborn 0.11.1 documentation (pydata.org)

Mục tiêu: Sinh viên nắm được các kiến thức sau

1. Data Cleansing
2. Exploration Data Analysis (EDA)
3. Kỹ thuật function chain trong Pandas – pipe()
4. Feature Engineering
5. Data Wrangling

Mô tả dữ liệu: The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren't enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew. While there was some

element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

Variable	Definition	Key
PassengerId	Identifier	
Survived	Survival	0 = No, 1 = Yes
Pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Name	Passenger name	
Sex	Sex	
Age	Age in years	
SibSp	# of siblings / spouses aboard the Titanic	
Parch	# of parents / children aboard the Titanic	
Ticket	Ticket number	
Fare	Passenger fare	
Cabin	Cabin number	
Embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Yêu cầu: *Hãy chuẩn bị dữ liệu phục vụ cho bài toán: "Xây dựng mô hình dự báo nhóm hành khách có khả năng sống sót với các thông số đầu vào là các đặc trưng của hành khách (name, age, gender, socio-economic class, ...), trong sự kiện Titanic lịch sử"*

PHẦN 1: DATA CLEANSING & FEATURE ENGINEERING

Hướng dẫn

1. Viết hàm load_data() để tải dữ liệu lên ứng dụng.Sau đó, hiển thị ra màn hình 10 dòng đầu tiên.
2. Thống kê dữ liệu thiếu trên các biến số và trực quan hóa dữ liệu thiếu bằng biểu đồ (Heat map). Hãy cho nhận xét về tình trạng thiếu dữ liệu Age, Cabin và Embarked
3. Xử lý tên cột tên Name, tách ra làm 2 cột: firstName và secondName. Lưu ý: Sau khi tách cột xong thì xóa luôn cột Name
4. Xứ lý rút gọn kích thước dữ liệu trên cột Sex như sau: thay thế male → M và female → F
5. Xử lý dữ liệu thiếu trên biến Age bằng cách thay thế bằng giá trị trung bình tuổi: Hãy đưa ra quyết định dùng giá trị trung bình tuổi toàn bộ hành khách hay theo từng nhóm hạng vé (hạng hành khách: Pclass). Ta tiến hành làm các bước sau
    a. Sử dụng Seaborn để vẽ biểu đồ (Box plot) trực quan dữ liệu để xác định phân phối tuổi trên từng hạng hành khách. Nhận xét về tuổi trung bình giữa các nhóm hành khách. Từ đó đưa ra quyết định cách thay thế giá trị tuổi bị thiếu.
    b. Tiến hành thay thế giá trị Age bị thiếu. Sau đó, hiển thị kết quả dạng bảng và trực quan dữ liệu đã xử lý thiếu cho cột'Age' bằng biểu đồ Heat map.
6. Xây dựng biến số Agegroup có thang đo thứ tự được ánh xạ theo thang đo khoảng dựa trên độ tuổi của hành khách như sau: (age =< 12] → Kid; (12, 18]: Teen, (18, 60]: Adult và (age > 60): Older
7. Tiến hành thêm đặc trưng về danh xưng (namePrefix) trong xã hội bằng cách tách Mr, Mrs, Miss, Master ra khỏi "secondName"

8. Khai thác thêm thông tin số lượng thành viên đi theo nhóm thân quen (familySize) đối với mỗi hành khách trên chuyến hải trình; family size = 1+ SibSp + Parch
9. Tạo thêm đặc trưng 'Alone' để xác định hành khách đi theo nhóm hay cá nhân bằng cách dựa trên familySize như sau: Nếu familySize = 0 thì giá trị Alone = 1 và ngược lại là 0.
10. Tiến hành tách loại cabin (typeCabin) mà hành khách ở để lọc và phân tích đặc tính cabin. Loại cabin được kí hiệu bởi chữ cái đầu tiên. Lưu ý: Đối với dữ liệu cabin bị thiếu thì thay thế bằng "Unknown"
11. Loại bỏ dữ liệu thừa đối với các hành khách xuất hiện trong cả 2 tập dữ liệu huấn luyện (train.csv) và đánh giá (test.csv). Ưu tiên giữ lại dữ liệu trong tập huấn luyện.

PHẦN 2: KHAI THÁC THÔNG TIN HỮU ÍCH – EDA

Hướng dẫn: ***Sinh viên cần đưa ra nhận xét sau mỗi biểu đồ trực quan nhằm rút trích được thông tin có giá trị về hành khách sống sót dựa trên các đặc trưng bên trên***

12. Trực quan thông tin tương quan tỉ lệ sống sót và thiệt mạng trên từng nhóm giới tính.
13. Trực quan thông tin hành khách sống sót trên từng nhóm phân loại hành khách (Pclass).
14. Trực quan thông tin hành khách sống sót trên từng nhóm giới tính và thang đo tuổi tác
15. Trực quan xác suất hành khách sống sót dựa trên thông tin nhóm đi cùng
16. Trực quan xác suất hành khách sống sót dựa trên thông tin giá vé
17. Trực quan số lượng người thiệt mạng và sống sót theo phân lớp (Pclass) hành khách và cảng sẽ cập bến.

LAB 3:

LÀM SẠCH DỮ LIỆU CƠ BẢN

Nội dung: Xử lý dữ liệu y khoa về huyết áp của bệnh nhân

Mục tiêu: Sinh viên biết cách sử dụng gói Pandas để xử lý dữ liệu

1. Tiến hành hiểu dữ liệu từ chuyên gia

   "*The data set has been kept small enough for you to be able to grok it all at once. The data is in csv format. Each row in the dataset has data about different individuals and their heart rate details for different time intervals. The columns contain information such as individual's Age, Weight, Sex and Heart Rates taken at different time intervals.*"
2. Thông thường ta thường xử lý các vấn đề sau về dữ liệu
   - Thiếu dòng tiêu đề ở file csv
   - Nhiều biến lưu ở một cột
   - Dữ liệu cột chứa các giá trị đơn vị không nhất quán
   - Dữ liệu có một dòng trống
   - Dữ liệu có các dòng trùng lặp
   - Các ký tự không phải ASCII
   - Giá trị bị mất
   - Tiêu đề cột là giá trị chứ không phải tên biến
3. ***Vấn đề 1:*** Tiến hành tải dữ liệu vào chương trình ứng dụng Python và giải quyết vấn đề "Missing header in the csv file"

```
#Problem 1
# Thêm header vào dataframe để diễn giải dữ liệu
column_names= ["Id", "Name", "Age", "Weight",'m0006','m0612','m1218','f0006','f0612','f1218']
# Đọc file dữ liệu
df = pd.read_csv("patient_heart_rate.csv", names = column_names)
#Hiển thị một vài dòng dữ liệu đầu tiên ra màn hình
print(df.head())
```

4. ***Vấn đề 2:*** Xử lý vấn đề một cột lưu hỗn hợp nhiều dữ liệu, ở đây là cột "Name" chứa bao gồm "Firstname" và "Lastname", giải pháp là ta sẽ tách ra làm 2 cột

```
#Problem 2
df[['Firstname','Lastname']] = df['Name'].str.split(expand=True)
df = df.drop('Name', axis=1)
print (df)
```

5. ***Vấn đề 3:*** Cột Weight có vấn đề về không thống nhất các đơn vị đo lường trong dữ liệu. Ta sẽ chuyển các đơn vị về thành đơn vị chuẩn "kg"

```
# Problem 3
#Get the Weight column
weight = df['Weight']

for i in range (0 ,len(weight)):
    x= str(weight[i])
    #Incase lbs is part of observation remove it
    if "lbs" in x[-3:]:
        #Remove the lbs from the value
        x = x[:-3:]
        #Convert string to float
        float_x = float(x)
        #Covert to kgs and store as int
        y =int(float_x/2.2)
        #Convert back to string
        y = str(y)+"kgs"
        weight[i]= y
print (df)
```

6. ***Vấn đề 4:*** Vấn đề về xuất hiện dòng dữ liệu rỗng (không có giá trị: NaN). Giải pháp có thể đưa ra là xóa bỏ

```
# Problem 4:
df.dropna(how="all", inplace=True)
print(df)
```

7. ***Vấn đề 5:*** Có nhiều dòng dữ liệu bị trùng lắp thông tin hoàn toàn[fullname, lastname, age, weight,....], giải pháp đưa ra là chỉ giữ lại một dòng dữ liệu, tuy nhiên giải pháp phải dựa trên nghiệp vụ của tập dữ liệu và quan sát của người xử lý.

```
df = df.drop_duplicates(subset=['Firstname','Lastname', 'Age', 'Weight'])
print (df)
```

8. ***Vấn đề 6:*** Xuất hiện dữ liệu bị ảnh hưởng bởi lỗi non-ASCII, không định dạng ASCII. Giải pháp: Tùy vào nghiệp vụ ta có thể: xóa dữ liệu tại đó, thay thế bằng dữ liệu khác hoặc thay bằng việc đánh dấu bằng một kí tự khác (ví dụ: 'warning')

```
#Problem 6:
df.Firstname.replace({r'[^\x00-\x7F]+':''}, regex=True, inplace=True)
df.Lastname.replace({r'[^\x00-\x7F]+':''}, regex=True, inplace=True)
print (df)
```

9. ***Vấn đề 7:*** "Missing values", vấn đề này xảy ra tại các cột "Age", "Weight" và "Heart Rate". Thiếu dữ liệu (dữ liệu không đầy đủ) là vấn đề xảy ra nhiều trong các nguồn dữ liệu do nhiều nguyên nhân chủ quan lẫn khách quan.Có một vài giải pháp để xử lý vấn đề này, chủ yếu dựa trên kinh nghiệm và nghiệp vụ về tập dữ liệu đó. Một số giải pháp đưa đề xuất từ chuyên gia như sau:
    a. **Deletion**: Remove records with missing values
    b. **Dummy substitution**: Replace missing values with a dummy but valid value: e.g.: 0 for numerical values.
    c. **Mean substitution**: Replace the missing values with the mean.
    d. **Frequent substitution**: Replace the missing values with the most frequent item.
    e. **Improve the data collector**: Your business folk will talk to the clients and inform them about why it is worth fixing the problem with the data collector.

    ……

    Yêu cầu:

    - Thống kê thông tin dữ liệu thiếu trên từng biến Age và Weight
    - Yêu cầu xử lý dữ liệu thiếu như sau: Nếu dòng nào có Age hoặc Weight có dữ liệu thì phần Age hoặc Weight được tính như bên dưới, nếu thiếu cả 2 thông tin thì xóa dòng
        - *Age*: Giá trị thay thế là mean của các giá trị trong cột Age
        - *Weight*: Giá trị thay thế là mean của các giá trị trong cột Weight theo nhóm giới tính.

10. ***Vấn đề 8:*** "một cột chứa quá nhiều thông tin cần được phân rã", như trong bài toán này ta thấy header "m0006" chứa các nội dung bao gồm: m → male, 1218 ~ 12-18 (mm-dd). Còn giá trị thì là kết quả huyết áp.

3	4.0	NaN	78kgs	78	79	72	-	-	-	Scrooge	McDuck
4	5.0	54.0	90kgs	-	-	-	69	NaN	75	Pink	Panther
5	6.0	52.0	85kgs	-	-	-	68	75	72	Huey	McDuck
6	7.0	19.0	56kgs	-	-	-	71	78	75	Dewey	McDuck
7	8.0	32.0	78kgs	78	76	75	-	-	-	Scööpy	Doo

Chúng ta sẽ tách nội dung của cột này ra làm 3 cột sau: PulseRate : giá trị huyết áp, Sex: giới tính (m: male, f: female) và time: thời gian (tháng-ngày) như sau:

	Id	Age	Weight	Firstname	Lastname	PulseRate	Sex	Time
0	1.0	56.0	70kgs	Micky	Mous	72	m	00-06
9	1.0	56.0	70kgs	Micky	Mous	69	m	06-12
18	1.0	56.0	70kgs	Micky	Mous	71	m	12-18
27	1.0	56.0	70kgs	Micky	Mous	-	f	00-06
36	1.0	56.0	70kgs	Micky	Mous	-	f	06-12
45	1.0	56.0	70kgs	Micky	Mous	-	f	12-18

Gợi ý:

```
#Melt the Sex + time range columns in single column
df = pd.melt(df,id_vars=['Id','Age','Weight','Firstname','Lastname'], value_name="PulseRate",var_name="sex_and_time").sort_values(['Id','Age','Weight','Firstname','Lastname'])

# Extract Sex, Hour Lower bound and Hour upper bound group
tmp_df = df["sex_and_time"].str.extract("(\D)(\d+)(\d{2})",expand=True)

# Name columns
tmp_df.columns = ["Sex", "hours_lower", "hours_upper"]

# Create Time column based on "hours_lower" and "hours_upper" columns
tmp_df["Time"] = tmp_df["hours_lower"] + "-" + tmp_df["hours_upper"]

# Merge
df = pd.concat([df, tmp_df], axis=1)

# Drop unnecessary columns and rows
df = df.drop(['sex_and_time','hours_lower','hours_upper'], axis=1)
df = df.dropna()
df.to_csv('outputcleanup.csv',index=False)
print (df)
```

11. Hãy khảo sát tỉ lệ dữ liệu thiếu trên biến huyết áp. Dữ liệu bị thiếu thì hãy xử lý bằng phương pháp sau
    - Thay thế bằng giá trị trung bình liền trước và liền sau của người đó. Nếu không được thì dùng 2)
    - Thay thế bằng giá trị trung bình 2 giá liền trước của người đó. Nếu không được thì dùng 3)
    - Thay thế bằng giá trị trung bình 2 giá liền sau của người đó. Nếu không được thì dùng 4)
    - Trung bình của các giá trị huyết áp của người đó. Nếu không được thì dùng 5).
    - Trung bình của các giá trị huyết áp của nhóm giới tính. Nếu không được thì dùng 6)
    - Trung bình của các giá trị dữ liệu. Nếu không được thì thay bằng mức ổn định trong y học.
12. Hãy rút gọn dữ liệu phù hợp và reindex lại dữ liệu. Sau đó, lưu trữ dữ liệu đã xử lý thành công với tên file ***patient_heart_rate_clean.csv***

Lưu ý: Ngoài ra còn rất nhiều vấn đề về mặt xử lý dữ liệu dựa trên nhiều khía cạnh khác nhau tùy vào sự am hiểu về dữ liệu của các chuyên gia như:

- Handling dates
- Correcting character encodings (a problem you hit when you scrape data off the web)

LAB 4:

BÀI THỰC HÀNH TRÌNH BÀY DỮ LIỆU

Nội dung: Trực quan hóa dữ liệu điểm thi đã được xử lý **processed_dulieuxettuyendaihoc.csv**
Mục tiêu: Sinh viên đạt được kiến thức sau.

1. Trình bày dữ liệu cơ bản
2. Trực quan hóa dữ liệu cơ bản

Phần 1: Thống kê dữ liệu

1. Hãy sắp xếp dữ liệu điểm DH1 theo thứ tự tăng dần

2. Hãy sắp xếp dữ liệu điểm DH2 tăng dần theo nhóm giới tính
3. Hãy tạo pivot-table để thống kê các giá trị count, sum, mean, median, min, max, sdt, Q1, Q2 và Q3 của DH1 theo KT
4. Hãy tạo pivot-table để thống kê các giá trị count, sum, mean, median, min, max, sdt, Q1, Q2 và Q3 của DH1 theo KT và KV
5. Hãy tạo pivot-table để thống kê các giá trị count, sum, mean, median, min, max, sdt, Q1, Q2 và Q3 của DH1 theo KT, KV và DT

Phần 2: Trình bày dữ liệu

1. Hãy trình bày dữ liệu biến: GT

Gợi ý

- Lập bảng tần số và tần suất
- Vẽ biểu đồ tần số (cột), biểu đồ tần suất (tròn) và biểu đồ tích lũy tần suất (đa giác tích lũy)

2. Hãy trình bày dữ liệu lần lượt các biến: **US_TBM1, US_TBM2 và US_TBM3**
3. Hãy trình bày dữ liệu biến DT với các học sinh là nam
4. Hãy trình bày dữ liệu biến KV với các học sinh là nam thuộc dân tộc Kinh, có điểm thõa mãn điều kiện (DH1 >= 5.0 và DH2 >= 4.0 và DH3 >= 4.0)
5. Hãy trình bày dữ liệu lần lượt các biến DH1, DH2, DH3 lớn hơn bằng 5.0 và thuộc khu vực 2NT

Phần 3: Trực quan hóa dữ liệu theo nhóm phân loại

1. Trực quan dữ liệu học sinh nữ trên các nhóm XL1, XL2, XL3 dạng unstacked

Gợi ý

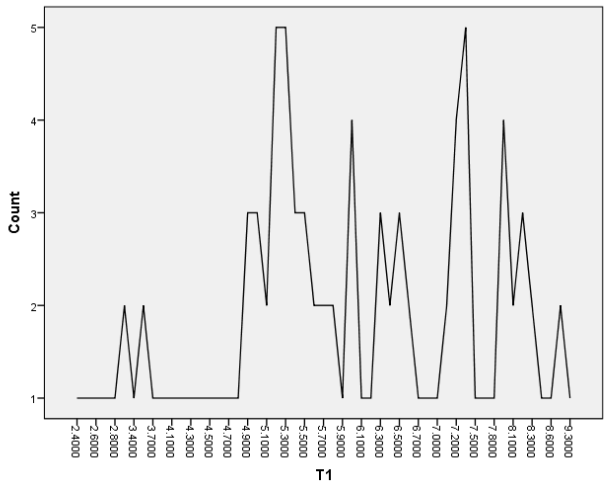
- Lọc dữ liệu giới tính là nữ
- Oy: Chiều cao biểu đồ cột thể hiển số lượng học sinh theo xếp loại
- Màu sắc thể hiện giá trị xếp loại: [Y, TB, K, G, XS]
- Ox: thể hiện nhóm XL1, XL2 và XL3

2. Trực quan dữ liệu KQXT trên nhóm học sinh có khối thi A, A1, B thuộc khu vực 1, 2
3. Trực quan dữ liệu số lượng thí sinh từng khu vực dựa trên từng nhóm khối thi
4. Trực quan dữ liệu số lượng thí sinh đậu, rớt trên từng nhóm khối thi
5. Trực quan dữ liệu số lượng thí sinh đậu rớt trên từng nhóm khu vực.
6. Trực quan dữ liệu số lượng thí sinh đậu rớt dựa trên từng nhóm dân tộc
7. Trực quan dữ liệu số lượng thí sinh đậu rớt dựa trên từng nhóm giới tính.

Phần 4: Trực quan hóa dữ liệu nâng cao

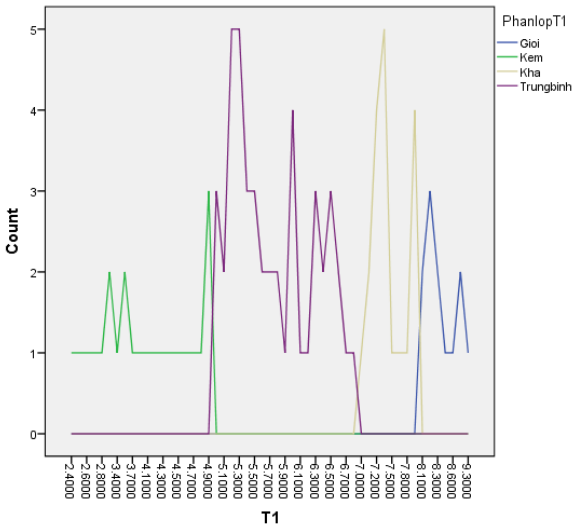
1. Vẽ biểu đồ đường Simple cho biến T1

Kết quả



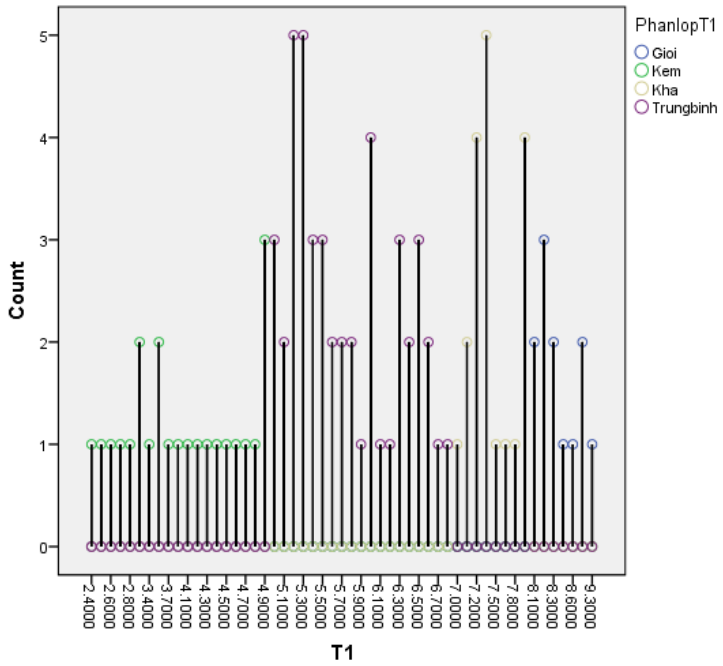
2. Hãy tạo biến phân loại (phanlopt1) cho môn toán (T1) như sau:
    a. Từ 0 đến dưới 5 = kém (ký hiệu "k")
    b. Từ 5 đến dưới 7 = trung bình (ký hiệu "tb")
    c. Từ 7 đến dưới 8 = khá (ký hiệu "k")
    d. Từ 8 trở lên = giỏi (ký hiệu "g")
3. Lập bảng tần số cho biến phanloait1
4. Vẽ biểu đồ đường Multiple Line cho biến T1 được phân loại bởi biến phanlopt1

Kết quả



5. Vẽ biểu đồ Drop-line cho biến T1 được phân loại bởi biến phanlopt1

Kết quả



Phần 5: Mô tả dữ liệu và khảo sát dạng phân phối

1. Hãy mô tả và khảo sát phân phối cho biến T1

Gợi ý

- Mô tả độ tập trung và phân tán của dữ liệu T1
- Vẽ biểu đồ Box-Plot và xác định các 10 đại lượng trong biểu đồ đó
- Mô tả hình dáng lệch của phân phối T1 dựa vào các đại lượng hướng tâm
- Vẽ biểu đồ Histogram biểu thị hình dáng phân phối
- Mô tả các đặc trưng của phân phối, mức độ lệch và mức độ nhọn
- Kiểm chứng phân phối chuẩn QQ-Plot
- Nhận xét và đánh giá về phân phối của T1

2. Hãy mô tả và khảo sát phân phối cho biến T1 trên từng nhóm phân lớp (phanlopT1)

Gợi ý

- Trực quan hóa biểu đồ Box-plot, histogram và QQ-plot theo phân nhóm là giá trị của 'phanlopT1'.

3. Hãy khảo sát tương quan giữa biến DH1 theo biến T1

Gợi ý

- Nhận xét giá trị Covariance hoặc Correlation
- Vẽ biểu đồ Scatter thể hiện liên hệ của biến phụ thuộc DH1 theo biến độc lập T1

4. Hãy khảo sát tương quan giữa biến DH1 theo biến T1 trên từng nhóm khu vực
5. Hãy khảo sát tương quan giữa các biến DH1, DH2, DH3

Gợi ý

- Nhận xét ma trận hiệp phương sai hoặc ma trận tương quan
- Vẽ biểu đồ Scatter giữa các biến

LAB 5:

BÀI TẬP TỔNG HỢP

Nội dung:

Trong lab này, chúng ta sẽ:
1. Học cách phân tích dữ liệu thông qua các giá trị tóm tắt dữ liệu và qua biểu diễn hình học của dữ liệu.
2. So sánh hai tập dữ liệu

Dữ liệu: Dữ liệu sử dụng trong lab này là tập dữ liệu về cân nặng của trẻ sơ sinh trong trường hợp bà mẹ hút thuốc lá khi mang thai và trong trường hợp bà mẹ không hút thuốc lá khi mang thai. (Dữ liệu được chuẩn bị sẵn trong tập tin: babies.txt).

Mô tả dữ liệu:

Tên cột	Ý nghĩa
bwt	Cân nặng của trẻ sơ sinh (baby weight), tính theo đơn vị ounce (100 ounce=2.83495kg)
smoke	Tình trạng hút thuốc của bà mẹ khi mang thai. 0= không hút, 1= có hút, 9=không biết

I. CÁC NỘI DUNG CẦN TÌM HIỂU:

Để thực hiện được lab này, sinh viên cần vận dụng các kiến thức ở các lab trên vào bài toán cụ thể:

1) Ước lượng độ biến động của dữ liệu:

Hai yếu tố chính để ước lượng độ biến động của dữ liệu: tâm và đuôi dữ liệu. Qua đó, ta cần tìm hiểu: dữ liệu phân bố như thế nào ở trung tâm (center) và như thế nào ở hai bên đuôi (tail).

Trong dữ liệu một chiều, để đo tính biến động của dữ liệu, ta có thể sử dụng các đại lượng: phương sai (Variance), độ lệch chuẩn (Standard deviation), khoảng cách giữa giá trị lớn nhất và nhỏ nhất (Range) và phần tư vị (IQR-InterQuantile Range). IQR cho phép khảo sát phần tâm dữ liệu trong khoảng từ ¼ cho đến ¾.

Đôi khi, để dễ hình dung, người phân tích có thể biểu diễn dữ liệu theo boxplot hay histogram, sẽ minh họa sau.

2) Phân tích về hình dạng của phân phối dữ liệu:

Để phân tích hình dạng phân phối dữ liệu, người phân tích cần tính giá trị **KURTOSIS**, là giá trị để đo độ "bè-nhọn" của đỉnh dữ liệu và giá trị **SKEWNESS** để đo độ "lệch (trái, phải)" của dữ liệu.

3) Phân tích tính chuẩn:

Để phân tích xem dữ liệu có phân phối chuẩn hay không, một cách trực quan, ta biểu diễn theo đường cong chuẩn (normal curve) và đôi khi cần một số thao tác chuẩn hóa.

II.CÁC NỘI DUNG THỰC HIỆN:

Trong lab này, ta phân tích các dữ liệu quan sát được để trả lời câu hỏi: "Việc bà mẹ hút thuốc khi mang thai có ảnh hưởng đến cân nặng của trẻ sơ sinh hay không?"
Để trả lời câu hỏi trên, cần thực hiện so sánh cân nặng của trẻ sơ sinh trong hai trường hợp: trường hợp bà mẹ hút thuốc khi mang thai và trường hợp bà mẹ không hút thuốc khi mang thai. Sự khác biệt đó có ý nghĩa hay không?
Để so sánh cân nặng của trẻ sơ sinh trong 2 trường hợp, có thể dựa vào thống kê mô tả: thống kê mô tả bằng số (numerical summaries), thống kê mô tả bằng hình (graphical): histogram, boxplot, quantile plot. Do đó, các nội dung chi tiết cần thực hiện:

1) Tính các đại lượng thống kê mô tả từ đó rút ra nhận xét về từng tập dữ liệu (cân nặng của trẻ trong trường hợp bà mẹ hút thuốc và cân nặng của trẻ trong trường hợp bà mẹ không hút thuốc).
   Cụ thể, ta sẽ phân tích sự khác biệt giữa hai tập dữ liệu: cân nặng của trẻ trong trường hợp bà mẹ hút thuốc và cân nặng của trẻ trong trường hợp bà mẹ không hút thuốc dựa vào các đại lượng thống kê mô tả.
2) Biểu diễn dữ liệu dưới các dạng đồ thị từ đó rút ra nhận xét về từng tập dữ liệu (trường hợp bà mẹ hút thuốc và trường hợp bà mẹ không hút thuốc)
   Cụ thể, ta sẽ sử dụng các dạng đồ thị: histogram, boxplot, quantile qua đó phân tích sự khác biệt giữa hai tập dữ liệu: cân nặng của trẻ trong trường hợp bà mẹ hút thuốc và cân nặng của trẻ trong trường hợp bà mẹ không hút thuốc dựa vào các đồ thị.

HƯỚNG DẪN THỰC HIỆN:

1.Mô tả dữ liệu bằng các giá trị số:

Bước 1: Tính các đại lượng thống kê cho hai tập dữ liệu:

(Cân nặng của trẻ trong trường hợp bà mẹ hút thuốc khi mang thai và cân nặng của trẻ trong trường hợp bà mẹ không hút thuốc khi mang thai).

Dùng python để thực hiện, kết quả được trình bày trong bảng sau:

	TH1: Bà mẹ hút thuốc	TH2: Bà mẹ không hút thuốc
Số lượng	484	742
Min	58	55
Max	163	176
Mean	114.10950413223141	123.04716981132076
Sd	18.09894568615237	17.39868877808027
Var	327.57183495029346	302.7143711964963

Median	115.0	123.0
Quantile 0%	58.0	55.0
Quanlite 25%	102.0	113.0
Quanlite 50%	115.0	123.0
Quantile 75%	126.0	134.0
IQR	24.0	21.0
Skewness	-0.03359497605204854	-0.18698408606617228
Kurtosis	2.988032478793404	4.037060312433822

Bước 2: Phân tích dữ liệu dựa trên các đại lượng vừa tính.

1. Xét tập dữ liệu ứng với trường hợp bà mẹ có hút thuốc

Vị trí tập trung của dữ liệu: khoảng giá trị: 114-115

Tính biến động của dữ liệu:

- **Phương sai (variance):** var= 327.57183495029346
- **Độ lệch chuẩn (standard deviation): sd=** 18.09894568615237
- **Khoảng giá trị:** min=58, max=163→ range=105
- **Khoảng cách giữa 2 phần tư vị:** IQR=Q3-Q1=126-102=24

Nhận xét: Như vậy dữ liệu phân bố gần nhau.

Hình dạng phân bố của dữ liệu:

- **Độ lệch:** Skewness=-0.03359497605204854
- **Độ bè nhọn của đỉnh dữ liệu:** Kurtosis=2.988032478793404

Nhận xét: Như vậy dữ liệu hơi lệch về phía trái, và đỉnh nhọn, hai bên giảm với tốc độ vừa phải.

2. Xét tập dữ liệu ứng với trường hợp bà mẹ không hút thuốc

Phần này sinh viên tự thực hiện.

Bước 3: So sánh các giá trị thống kê mô tả của hai tập dữ liệu.

Sự khác biệt về vị trí tập trung dữ liệu: chênh lệch khoảng 123 -115 = 8

Nhận xét: khác biệt không đáng kể.

Sự khác biệt về tính biến động của dữ liệu được thể hiện qua bảng sau:

	TH1: Bà mẹ hút thuốc	TH2: Bà mẹ không hút thuốc	Chênh lệch (TH2-TH1)
Sd	18.09894568615237	17.39868877808027	-0.700256908
Var	327.57183495029346	302.7143711964963	-24.85746375
Range	163-58=105	176-55=121	16
IQR	126-102=24	134-113=21	-3

Dữ liệu trong trường hợp bà mẹ không hút thuốc có phân bố rộng hơn nhưng phần dữ liệu tập trung lại hẹp hơn so với trường hợp bà mẹ có hút thuốc. Sự biến động của dữ liệu trong hai trường hợp không khác biệt nhiều.

Sự khác biệt về hình dạng phân bố của dữ liệu: được thể hiện qua bảng sau:

	TH1: Bà mẹ hút thuốc	TH2: Bà mẹ không hút thuốc	Chênh lệch (TH2-TH1)
Skewness	-0.03359497605204854	-0.18698408606617228	-0.15338911

Kurtosis	2.988032478793404	4.037060312433822	1.049027834

- **Nhận xét:** trường hợp bà mẹ hút thuốc có phân bố dữ liệu nhọn hơn, đối xứng hơn so với trường hợp không hút thuốc. Cả 2 trường hợp đều hơi lệch về trái.

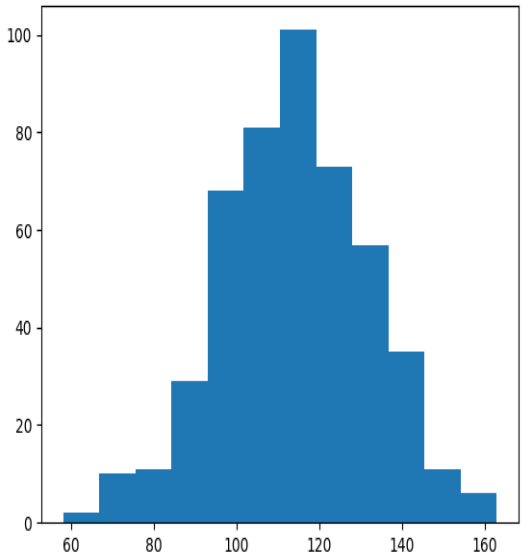
2. Biểu diễn hình học của dữ liệu

1. Dữ liệu cân nặng của trẻ trong trường hợp bà mẹ hút thuốc và bà mẹ không hút thuốc

Ta sẽ phân tích các biểu đồ:

- Histogram
- Boxplot

a) Histogram trong trường hợp bà mẹ có hút thuốc:



Vị trí tập trung dữ liệu: khoảng 110

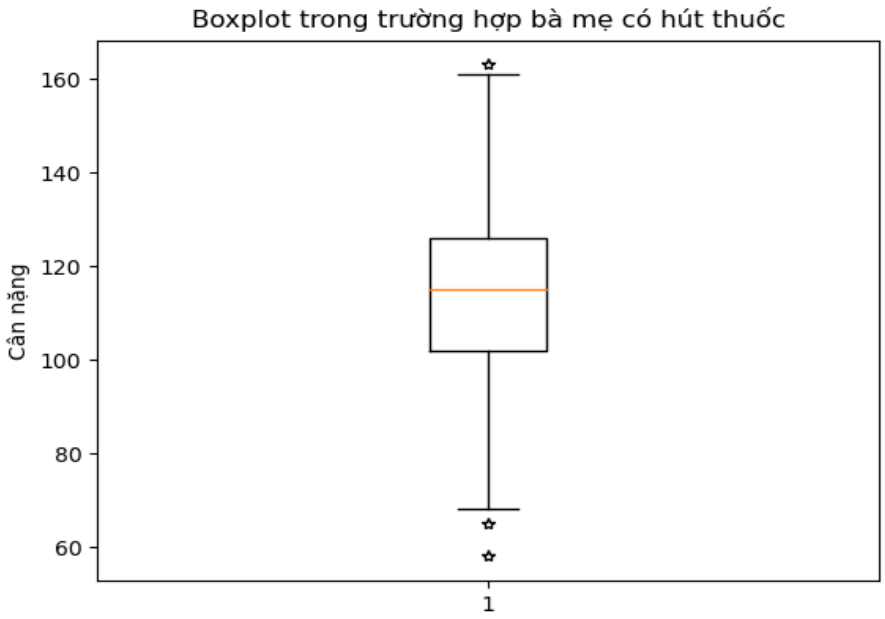
Tính biến động của dữ liệu: dữ liệu phân bố trong khoảng [50-170]

Tính đối xứng của phân bố dữ liệu: dữ liệu chỉ có 1 đỉnh. Bắt đầu từ đỉnh, hai bên giảm dần và tốc độ giảm vừa phải.

Dữ liệu phân bố gần đối xứng, hơi lệch về phía trái. Hai bên đuôi có độ dài vừa phải. Hai bên đỉnh dữ liệu cũng phân bố vừa phải.

Giá trị ngoại lệ: không thấy rõ có giá trị ngoại lệ nào đáng kể

b) Boxplot:



Tính biến động của dữ liệu: dữ liệu phân bố tập trung trong khoảng từ [102,126]

Giá trị ngoại lệ: có một số giá trị ngoại lệ (lớn hơn 162, nhỏ hơn 66) nhưng không nhiều.

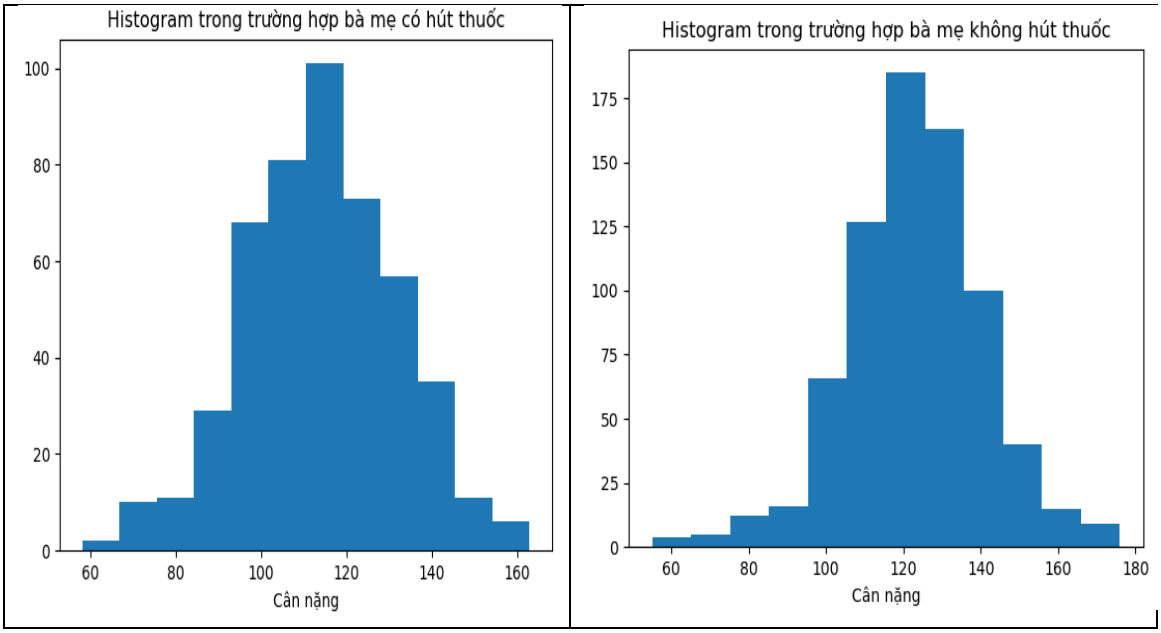
2. Dữ liệu cân nặng của trẻ trong trường hợp bà mẹ hút thuốc không hút thuốc

Phần này sinh viên tự thực hiện

So sánh hai tập dữ liệu dựa vào các biểu diễn hình học:

a) Histogram

Để so sánh, ta vẽ 2 histogram gần nhau:



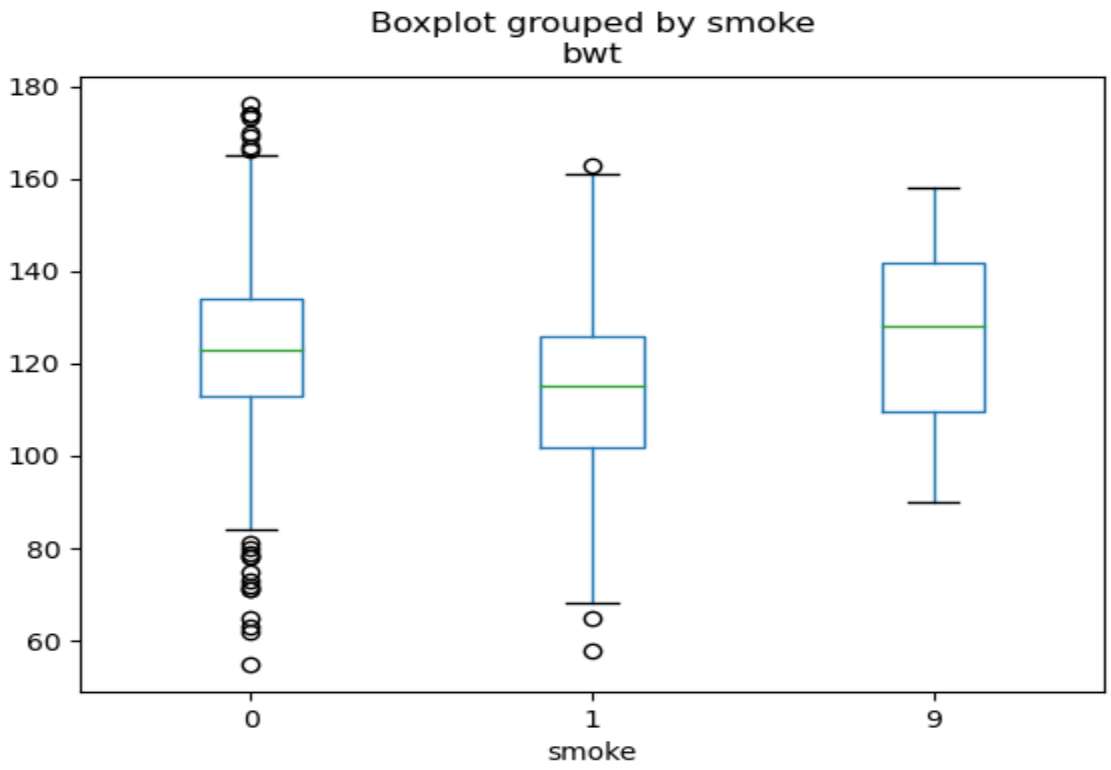
Cân nặng của trẻ trẻ trong trường hợp bà mẹ không hút thuốc cáo hơn so với trường hợp bà mẹ có hút thuốc

Tính biến thiên của 2 tập dữ liệu: tương tự nhau

Tính đối xứng của 2 tập dữ liệu: tương tự nhau

Giá trị ngoại lệ: cả 2 đều không có giá trị ngoại lệ đáng chú ý.

b) Boxplot



Khác biệt về vị trí: giá trị trung vị của trường hợp bà mẹ không hút thuốc lớn hơn trường hợp bà mẹ có hút thuốc (123 và 115). (Trường hợp smoke=9 là trường hợp không biết bà mẹ có hút thuốc hay không, trung vị trong trường hợp này cao hơn so với 2 trường hợp bà mẹ có hút thuốc và không hút thuốc).

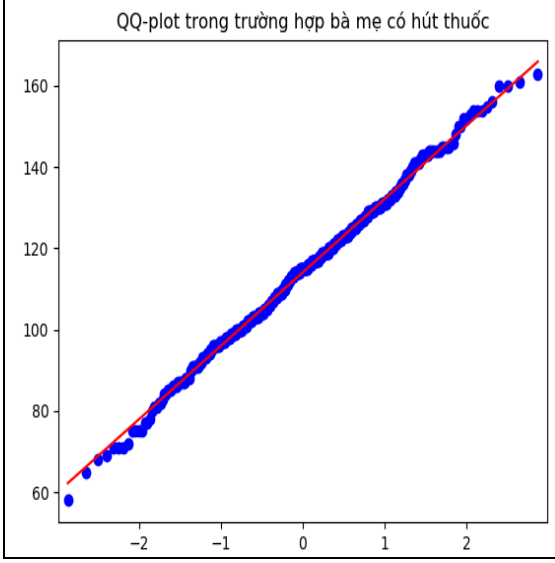
Giá trị ngoại lệ: cả 2 trường hợp đều có giá trị ngoại lệ trên và dưới. Trường hợp không hút thuốc có nhiều giá trị ngoại lệ hơn.

Ta dùng thêm đồ thị QQ-plot để phân tích

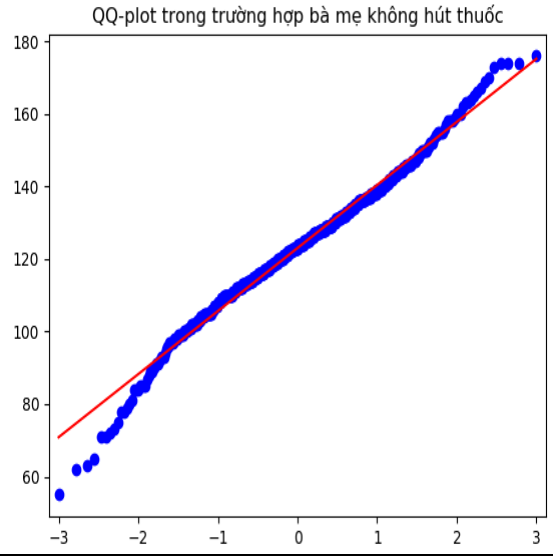
So sánh QQ-plot phân bố của 2 tập dữ liệu:

QQ-plot

Có hút thuốc



Không hút thuốc



QQ-plot có dạng đường thẳng, suy ra dữ liệu của 2 trường hợp có phân bố tương tự nhau.