Chapter 6

Chapter 6 - Exercise 1: Trực quan hóa dữ liệu Chipotle

Cho dữ liệu
'https://raw.githubusercontent.com/justmarkham/DAT8/master/data/chipotle.tsv'

Nhà hàng Chipotle cần phân tích dữ liệu bán được trong ngày diễn ra khuyến mãi để có thể điều chỉnh thực đơn và thực hiện các chương trình khuyến mãi phù hợp.

Dữ liệu được cung cấp trong file chipotle.tsv, hãy thực hiện các yêu cầu sau:

1. Đọc dữ liệu và gán vào biến chipo. Hiển thị 10 dòng đầu của dữ liệu.
2. Tạo biến x chứa các item_name là các món ăn được khách hàng gọi , in head của x
   Tạo một dictionary với 2 thông tin: tên món ăn (item_name) và tần suất/số lần gọi món (gợi ý: sử dụng collections.Counter(x)), in kết quả
3. Chuyển dictionary câu 2 thành DataFrame df để chuẩn bị cho các yêu cầu phân tích sau đó
4. a) Sắp xếp df theo tần suất giảm dần, và lấy 5 item đầu tiên
   b) Vẽ biểu đồ bar chart cho biết 5 món được gọi nhiều nhất (có title, xlabel, ylabel và xsticks)
5. a) Đổi kiểu dữ liệu của cột item_price sang kiểu số thực
   b) Nhóm các đơn hàng theo order_id, và tính tổng số lượng gọi và tổng giá trị của mỗi đơn hàng, in kết quả
6. Từ câu 5b, hãy vẽ scatterplot với x là item_price, và y là quantity, có title, xlabel, ylabel.
   Bạn có nhận xét gì qua biểu đồ này.

Chapter 6 - Exercise 2: Phân tích dữ liệu thế giới qua các năm

Cho các dữ liệu *year, pop, gdp_cap, life_exp, pop2, col* từ tập tin *data_year_pop_cap_life.txt*

Thực hiện các yêu cầu sau:

1. In item cuối của year và pop
2. Vẽ biểu đồ line thể hiện sự thay đổi dân số thế giới qua các năm (x-axis: year, y-axis: pop)

3. Cho biết thu nhập bình quân đầu người và tuổi thọ trung bình của item cuối trong gdp_cap và life_exp
4. Thử vẽ biểu đồ line liên hệ giữa gdp_cap và life_exp với x-axis: gdp_cap, y-axis: life_exp

   Biểu đồ này có thể xem được không? Nếu không thì bạn hãy đề xuất một loại biểu đồ phù hợp?
5. Vẽ biểu đồ histogram của life_exp, màu cột xanh, viền đỏ

   Bạn nhận xét gì qua biểu đồ vừa vẽ.
6. Vẽ biểu đồ histogram của life_exp, màu cột xanh dương, viền đỏ, với bins = 5, 15, 20 . Bạn nhận xét gì qua các biểu đồ vừa vẽ ?
7. Tạo scatter plot của gdp_gap và life_exp nhưng sử dụng plt.xscale('log').

   Khi trực quan hóa dữ liệu thay đổi trong phạm vi rất rộng, thang đo logarit plt.xscale('log') cho phép chúng ta hình dung các thay đổi một cách trực quan hơn.
8. Tạo scatter plot của gdp_gap và life_exp, sử dụng plt.xscale('log'). Thiết lập xlabel, ylabel, title

   Với: *tick_val = [1000,10000,100000]* và *tick_lab = ['1k','10k','100k']* => *plt.xticks(tick_val, tick_lab)*
9. Tạo numpy array np_pop từ pop2 trong file dữ liệu.

   Vẽ scatter plot của gdp_cap và life_exp, với s = np_pop * 2, màu magenta

   Thiết lập xlabel, ylabel, title và plt.xticks([1000, 10000, 100000],['1k', '10k', '100k'])
10. Danh sách màu tương ứng với các nhóm quốc gia: *'Asia':'red', 'Europe':'green', 'Africa':'blue', 'Americas':'yellow', 'Oceania':'black'*

    Vẽ scatter plot của gdp_cap và life_exp, với s = np.array(pop) * 2, màu c = col (giá trị col trong file dữ liệu), alpha=0.8

    Thiết lập xlabel, ylabel, title và plt.xticks([1000, 10000, 100000],['1k', '10k', '100k']).

    Bạn nhận xét gì về biểu đồ vừa vẽ ?
11. Vẽ scatter plot của gdp_cap, life_exp, với s = np.array(pop) * 2, màu c = col, alpha=0.8

    Thiết lập xlabel, ylabel, title và plt.xticks([1000, 10000, 100000],['1k', '10k', '100k'])

    Thêm text cho 2 nơi là India và China: plt.text(1550, 71, 'India'), plt.text(5700, 80, 'China')

    Thêm lưới cho biểu đồ.

Chapter 6 - Exercise 3: Titanic Disaster

Vào ngày 15 tháng 4 năm 1912, trong chuyến hành trình đầu tiên của mình, tàu Titanic đã gặp tai nạn sau khi va chạm với một tảng băng trôi, đã có 1502 người mãi mãi ra đi trong tổng số 2224 hành khách và phi hành đoàn.

Thông tin về Titanic Disaster có thể xem tại: https://www.kaggle.com/c/titanic/data

Dựa trên tập tin train.csv, hãy thực hiện các yêu cầu sau:

1. a) Đọc dữ liệu từ tập tin train.csv và lưu vào biến titanic. Hiển thị 5 dòng dữ liệu đầu của titanic.
   b) Thiết lập cột index cho titanic là PassengerId. Hiển thị lại 5 dòng dữ liệu đầu của titanic lúc này.
2. Tạo pie chart thể hiện tỷ lệ hành khách nam/nữ trên tàu
3. Cho biết có bao nhiêu người còn sống sót
4. Vẽ biểu đồ histogram của cột vé (Fare)
   Bạn nhận xét gì về biểu đồ vừa vẽ.

Chapter 6 - Exercise 4: Women in Science

Cho các dữ liệu *year, physical_sciences, computer_science, health, education* từ tập tin *women_in_science.txt*

Thực hiện các yêu cầu sau:

1. Vẽ biểu đồ line plot thể hiện tỷ lệ % bằng Khoa học vật lý và Khoa học máy tính được trao cho phụ nữ qua các năm
   Đồ thị có 2 line:
   - line 1: year, physical_sciences, color='blue';
   - line 2: year,computer_science, color='red'

   Bạn nhận xét gì về biểu đồ vừa vẽ ?
2. Vẽ 2 biểu đồ line plot ở câu 1 nhưng trên 2 vùng:
   - vùng 1: plt.axes([0.05, 0.05, 0.425, 0.9]),
   - vùng 2: plt.axes([0.525, 0.05, 0.425, 0.9])
3. Vẽ 2 biểu đồ line plot ở câu 1 nhưng trên 2 subplot:
   - subplot 1: plt.subplot(1, 2, 1),
   - subplot 2: plt.subplot(1, 2, 2)

Lưu ý: sử dụng plt.tight_layout() trước khi show()

4. Vẽ 4 biểu đồ line với 4 màu khác nhau:
    - Biểu đồ 1: thể hiện tỷ lệ % bằng Khoa học vật lý được trao cho phụ nữ qua các năm (year - physical_sciences)
    - Biểu đồ 2: thể hiện tỷ lệ % bằng Khoa học máy tính được trao cho phụ nữ qua các năm (year - computer_science)
    - Biểu đồ 3: thể hiện tỷ lệ % phụ nữ tham gia các công việc liên quan đến y tế qua các năm (year - health)
    - Biểu đồ 4: thể hiện tỷ lệ % phụ nữ tham gia các công việc liên quan đến giáo dục qua các năm (year - education)

    trên 4 subplot: plt.subplot(2, 2, 1), plt.subplot(2, 2, 2), plt.subplot(2, 2, 3), plt.subplot(2, 2, 4)

    Bạn nhận xét gì về biểu đồ vừa vẽ ?
5. Vẽ 2 biểu đồ line plot:
    - Biểu đồ 1: thể hiện tỷ lệ % bằng Khoa học máy tính được trao cho phụ nữ qua các năm (year - computer_science)
    - Biểu đồ 2: thể hiện tỷ lệ % bằng Khoa học vật lý được trao cho phụ nữ qua các năm (year - physical_sciences)

    nhưng giới hạn từ năm 1980-2000 --> có plt.xlim(1980, 2000) và plt.ylim(0, 50)

    Lưu biểu đồ này thành file hình.png
6. Vẽ 2 biểu đồ line plot:
    - Biểu đồ 1: thể hiện tỷ lệ % bằng Khoa học máy tính được trao cho phụ nữ qua các năm (year - computer_science)
    - Biểu đồ 2: thể hiện tỷ lệ % bằng Khoa học vật lý được trao cho phụ nữ qua các năm (year - physical_sciences)

    nhưng giới hạn từ năm 1990-2000 --> có plt.xlim(1990, 2000) và plt.ylim(0, 50)

    Bạn nhận xét gì về biểu đồ vừa vẽ ?

Chapter 6 - Exercise 5: WordClouds

Part 1:

- Cho dữ liệu Youtube04-Eminem.csv là dữ liệu được lấy từ UCI Machine Learning Repository, trong đó chứa các YouTube comment cho các video của các nghệ sỹ nổi tiếng.
- Hãy bỏ các STOPWORD
- Vẽ WordClouds cho dữ liệu này

Part 2:

- Sử dụng vietnamese stop word từ : https://github.com/stopwords/vietnamese-stopwords => tạo thành set các stop words
- Đọc file ngon_tu_quang_cao.txt => đưa nội dung vào biến text
- Bổ sung thêm một số từ không quan trọng vào stopwords
- Vẽ wordclouds
- Chọn hình làm wc_mask phù hợp => vẽ wordclouds với wc_mask

Chapter 6 - Exercise 6: TreeMap - Waffle Chart

Part 1: TreeMap

- Cho dữ liệu là danh sách ứng viên và số phiếu bầu trong cuộc bầu cử tổng thống Mỹ năm 2016.

	Hillary Clinton	Donald Trump	Others
Virginia	1.981.473	1.769.443	233.715
Maryland	1.677.928	943.169	160.349
West Virginia	188.794	489.371	36.258

- Vẽ 3 TreeMap thể hiện tỷ lệ số phiếu bầu lần lượt cho ứng viên ở Virginia, Maryland và West Virginia

Part 2: Waffle Chart

- Tính tổng số phiếu bầu của từng ứng viên ở cả 3 khu vực
- Vẽ Waffle Chart thể hiện tỷ lệ số phiếu bầu tổng cho từng ứng viên

Chapter 6 - Exercise 7: Area plot, Boxplot

Thực hành vẽ Area plot, Box plot trên 2 tập dữ liệu khác nhau.

Part 1: Area Plot

- Cho Dữ liệu số giờ nắng từ tháng 1 đến tháng 12 trong năm 2016, 2017 tại trạm quan trắc Vũng Tàu. Hiển thị nội dung của df.
  'Hours_2017': [183.4, 211.8, 286.4, 287.5, 238.8, 200.3, 187.4, 233.8, 225.5, 149.1, 180.2,198.3],
  'Hours_2016': [272.8, 254.0, 296.0, 298.0, 240.1, 197.8, 240.3, 219.5, 212.7, 134.7, 215.3, 109.1]
- Trên cùng một biểu đồ, hãy vẽ:
  - Area plot cho 12 tháng nắng trong năm 2016
  - Line plot cho 12 tháng nắng trong năm 2017
- Bạn nhận xét gì về biểu đồ vừa vẽ ?

Part 2: Boxplot

- Cho dữ liệu baseball.csv
  - Đọc dữ liệu từ baseball.csv và lưu vào biến data,
  - In 10 dòng nội dung đầu của data,
  - Cho biết thông tin thống kê chung của data
- Vẽ boxplot cho dữ liệu height và weight
- Kiểm tra xem dữ liệu có outliers hay không? Nếu có thì loại bỏ các outliers. Vẽ lại boxplot