BÀI 7: Thư viện Pandas – Xử lý dữ liệu

Mục tiêu chính: Cung cấp cho học viên kiến thức và kỹ năng sử dụng:

- *Thư viện Pandas*

7.1. Drinks

✓ **Dữ liệu**: Tập dữ liệu về tình hình tiêu thụ rượu bia ở các quốc gia theo từng châu lục. Bộ dữ liệu cung cấp thông tin về số lượng tiêu thụ bia, rượu mạnh, rượu vang theo từng châu lục (Dữ liệu **drinks.csv** được đính kèm trong file **data**).

✓ **Yêu cầu:**

1. Đọc dữ liệu từ tập tin **drinks.csv** với index_col là cột đầu tiên của dữ liệu, và lưu vào biến drinks
    a. Cho biết kích thước (shape) của **drinks**
    b. Hiển thị tên các cột (columns) của **drinks**
    c. Xem thông tin info()
    d. Xem 5 dòng dữ liệu đầu tiên (head) của **drinks**
    e. Xem 5 dòng dữ liệu cuối cùng (tail) của **drinks**
2. Đếm dữ liệu NULL

a. Hiển thị các dòng null

b. Xem xét thấy đây đều là các quốc gia ở Bắc Mỹ (continent: NA) nhưng bị hiểu nhầm là NaN, thay thế các giá trị NaN thành 'NA'

c. Cho biết số lượng quốc gia ở mỗi châu lục

3. Cho biết số lượng bia tiêu thụ trung bình (mean) ở mỗi châu lục

a. Sắp xếp giảm dần theo lượng bia tiêu thụ trung bình ở mỗi châu lục

b. Vẽ biểu đồ cột (Bar chart) với số lượng bia tiêu thụ trung bình ở mỗi châu lục

4. Cho biết thông tin thống kê tổng quát (describe) số lượng rượu vang được tiêu thụ ở mỗi châu lục

5. Cho biết số lượng bia và rượu tiêu thụ trung bình (mean) ở mỗi châu lục

6. Cho biết giá trị trung vị (median) của các loại bia và rượu tiêu thụ ở mỗi châu lục

7. Cho biết số lượng rượu mạnh (spirit_servings) tiêu thụ trung bình, lớn nhất và nhỏ nhất ở mỗi châu lục, chỉ tính ở các quốc gia có tiêu thụ rượu mạnh

8. Sắp xếp dữ liệu tăng dần (sort_values) theo số lượng bia tiêu thụ, chỉ tính ở các quốc gia có tiêu thụ bia

a. Cho biết 5 quốc gia có lượng tiêu thụ bia ít nhất

b. Cho biết 5 quốc gia có lượng tiêu thụ bia lớn nhất

c. Vẽ biểu đồ cột (bar chart) với 5 quốc gia có lượng tiêu thụ bia lớn nhất

9. Cho biết 5 quốc gia ở Châu Á có lượng tiêu thụ bia ít nhất

7.2. Pizza

✓ **Dữ liệu**: Tập dữ liệu về tình hình kinh doanh pizza của cửa hàng từ 01/01/2015 đến 08/01/2015. Bộ dữ liệu bao gồm 1000 dòng dữ liệu, cung cấp các thuộc tính như mã hóa đơn (order_id), tên bánh pizza (pizza_name), ngày đặt bánh (order_date) … và nhiều thông tin khác (Dữ liệu **Pizza_sales.xlsx** được đính kèm trong file **data**).

✓ **Yêu cầu**:

1. Đọc thông tin và xem thông tin dữ liệu
    a. Đọc tập tin **Pizza_sales.xlsx** vào dataframe có tên là data
    b. Xem thông tin: head(), tail(), info()
2. Chuyển đổi dữ liệu: chuyển đổi dữ liệu ở 2 cột **order_date** và **order_time** thành kiểu dữ liệu datetime
3. Tạo cột dữ liệu mới:
    a. Tạo thêm cột **order_day** và **order_hour** từ **order_date** và **order_time**
    b. Sắp xếp lại vị trí các cột dữ liệu
4. Thống kê:
    a. Số lượng bánh (**quantity**) khách hàng thường hay đặt
    b. Size bánh (**pizza_size**) khách hàng thường hay đặt
    c. Loại pizza (**pizza_category**) khách hàng thường hay đặt
    d. Top 10 hóa đơn có tổng thành tiền cao nhất và tính tổng số lượng bánh trên mỗi hóa đơn
    e. Cho biết thành tiền nhỏ nhất, lớn nhất và trung bình của các hóa đơn
    f. Top 10 món pizza bán chạy nhất
    g. Thứ mấy trong tuần là ngày bán chạy nhất?
    h. Khung giờ nào khách hàng đặt pizza nhiều nhất?

7.3. Điểm thi THPTQG 2016

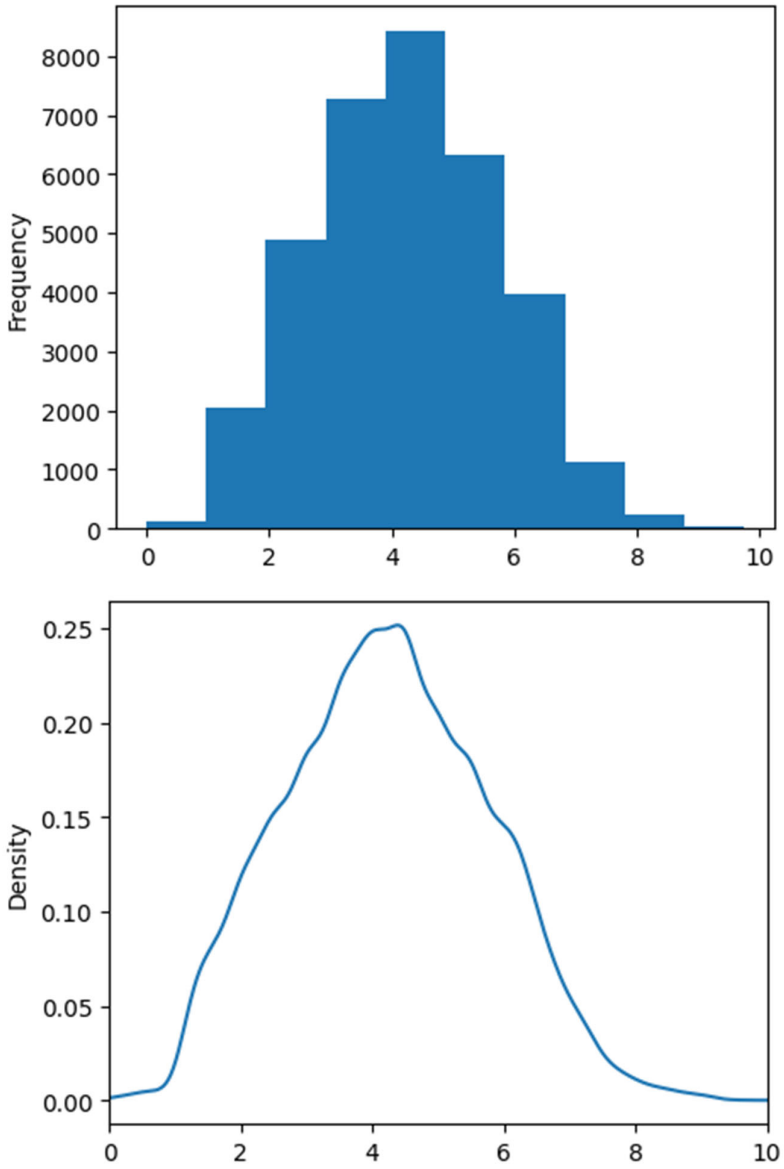
- ✓ **Dữ liệu**: Tập dữ liệu **Diemthi_thpt_quocgia_2016** chứa dữ liệu điểm thi THPT Quốc gia năm 2016 của gần 35.000 thí sinh. Danh sách các môn thi là: **'Toán', 'Ngữ văn', 'Địa lí', 'Lịch sử', 'Tiếng Anh', 'Sinh học', 'Vật lí', 'Hóa học'**. Một thí sinh chỉ thi các môn bắt buộc chung còn các môn tự chọn có thể khác nhau (Dữ liệu **Diemthi_thpt_quocgia_2016.csv** được đính kèm trong file **data**).

- ✓ **Yêu cầu**:

1. Đọc dữ liệu, xem thông tin dữ liệu
    a. Xem 5 dòng dữ liệu đầu tiên (head)
    b. Xem 5 dòng dữ liệu cuối cùng (tail)
    c. Xem thông tin info()
    d. Hiển thị tên các cột (columns)
2. Với dữ liệu hiện tại, cột **DIEM_THI** là chuỗi chứa điểm thi của tất cả các môn thi của một thí sinh
    a. Như vậy chúng ta sẽ không phân tích được điểm thi của các thí sinh. Do đó, việc đầu tiên là cần phải tiền xử lý dữ liệu
    b. Từ dữ liệu cột **DIEM_THI**, hãy tạo ra các cột tương ứng với danh sách các môn thi nói trên và đưa điểm của thí sinh từ chuỗi vào các cột, môn nào thí sinh không thi thì sẽ để NaN

3. Chuyển cột **NGAY_SINH** sang kiểu dữ liệu datetime

4. Hãy vẽ biểu đồ phân phối tần suất điểm thi, mỗi điểm thi là 1 biểu đồ, nhận xét trên từng biểu đồ



Biểu đồ phân phối tần suất điểm thi môn Toán

5. Lưu lại dữ liệu sau khi đã chuẩn hóa