Chapter 5

Chapter 5 - Exercise 1: GroupBy

drinks.csv là tập tin cung cấp dữ liệu về tình hình tiêu thụ rượu bia ở các quốc gia theo từng châu lục

1. Đọc dữ liệu từ tập tin drinks.csv với index_col là cột đầu tiên của dữ liệu, và lưu vào biến drink.
   Cho biết kiểu dữ liệu (type), kích thước (shape) của drink,
   Hiển thị tên các cột (columns) của drink,
   Xem 5 dòng dữ liệu đầu tiên (head) và cuối cùng (tail) của drink
2. Cho biết số lượng bia tiêu thụ trung bình ở mỗi châu lục
3. Cho biết thông tin thống kê tổng quát (describe) số lượng rượu vang được tiêu thụ ở mỗi châu lục
4. Cho biết số lượng các loại bia và rượu tiêu thụ trung bình (mean) ở mỗi châu lục
5. Cho biết giá trị trung vị (median) cho các loại bia và rượu tiêu thụ ở mỗi châu lục
6. Cho biết số lượng rượu mạnh (spirit_servings) tiêu thụ trung bình, lớn nhất và nhỏ nhất ở mỗi châu lục
7. Sắp xếp dữ liệu tăng dần (sort_values) theo số lượng bia tiêu thụ
   Cho biết 5 quốc gia có lượng tiêu thụ bia nhiều nhất,
   Cho biết 5 quốc gia có lượng tiêu thụ bia ít nhất

Chapter 5 - Exercise 2: Giao dịch chứng khoán

Cho 3 file .csv sau:

- *stocks1.csv : date, symbol, open, high, low, close, volume : chứa thông tin giao dịch chứng khoán các công ty khác nhau*
- *stocks2.csv : date, symbol, open, high, low, close, volume : chứa thông tin giao dịch chứng khoán các công ty khác nhau*
- *companies.csv : name, employees, headquarters_city, headquarters_state : chứa thông tin về trụ sở và số lượng nhân viên cho một công ty cụ thể*

Yêu cầu:

1. a) Đọc file stocks1.csv => đưa dữ liệu vào stocks1

Hiển thị 5 dòng dữ liệu đầu và cuối của stocks1
Cho biết kiểu dữ liệu (dtype) của các cột của stocks1
Xem thông tin (info) của stocks1
b) Đọc file stocks2.csv => đưa dữ liệu vào stocks2
Hiển thị 5 dòng dữ liệu đầu và cuối của stocks2
Cho biết kiểu dữ liệu (dtype) của các cột của stocks2
Xem thông tin (info) của stocks2
c) Đọc file companies.csv => đưa dữ liệu vào companies
Xem dữ liệu của companies
Cho biết kiểu dữ liệu (dtype) của các cột của companies
Xem thông tin (info) của companies

2. Cho biết trong stocks1 có dữ liệu Null hay không? Nếu có, hãy thay thế với quy tắc sau:
   - Nếu Null cột *high* thì thay bằng giá trị max trên cột high của mã chứng khoán đó.
   - Nếu Null cột *low* thì thay bằng giá trị min trên cột *low* của mã chứng khoán đó.
3. Tạo dataframe stocks bằng cách gộp stocks1 và stocks2 theo dòng. Xem 15 dòng dữ liệu cuối của stocks.
4. Tạo dataframe stocks_companies bằng cách gộp stocks và companies.
   Xem 5 dòng dữ liệu đầu của stocks_companies.
5. Cho biết giá (*open, high, low, close)* trung bình và *volume* trung bình của mỗi công ty.
6. Cho biết giá đóng cửa (close) trung bình, lớn nhất và nhỏ nhất ở mỗi công ty
7. Tạo cột parsed_time trong stocks_companies bằng cách đổi thời gian sang định dạng DateTime. Cho biết kiểu dữ liệu của cột parsed_time. Hiển thị 5 dòng dữ liệu đầu của stocks_companies
8. Thêm cột result, nếu giá 'close' > 'open' thì cột result có giá trị 'up', ngược lại 'down'.

Chapter 5 - Exercise 3: Phân tích dữ liệu Movies

Dữ liệu được lấy từ MovieLens website. Download the Dataset theo link:

- Data Source: MovieLens web site (filename: ml-latest-small.zip)
- Location: https://grouplens.org/datasets/movielens/latest/

Part 1: Đọc dữ liệu & Data Structures

Trong ml-latest-small.zip bao gồm 3 file CSV sau:

- *ratings.csv : userId,movieId,rating, timestamp : Chứa dữ liệu về các xếp hạng của các bộ phim, mỗi dòng biểu thị một xếp hạng của một phim bởi một người dùng.*
- *tags.csv : userId,movieId, tag, timestamp : chứa thông tin về các Tag mà người dùng gắn vào cho phim, mỗi dòng biểu thị cho 1 tag của một người dùng cho một phim*
- *movies.csv : movieId, title, genres : chứa thông tin về các bộ phim, mỗi dòng mô tả thông tin của 1 bộ phim*

Sử dụng pd.read_csv() để đọc dữ liệu

1. Đọc file movies.csv => đưa dữ liệu vào movies
   Cho biết kiểu dữ liệu (type), kích thước (shape) của movies
   Hiển thị 5 dòng dữ liệu đầu tiên (head) và cuối (tail) của movies
   Cho biết kiểu dữ liệu (dtype) của các cột của movies
   Xem thông tin (info) của movies
2. Đọc file tags.csv => đưa dữ liệu vào tags
   Cho biết kiểu dữ liệu (type), kích thước (shape) của tags
   Hiển thị 5 dòng dữ liệu đầu tiên (head) và cuối (tail) của tags
   Cho biết kiểu dữ liệu (dtype) của các cột của tags
   Xem thông tin (info) của tags
3. Đọc file ratings.csv => đưa dữ liệu vào ratings
   Cho biết kiểu dữ liệu (type), kích thước (shape) của ratings
   Hiển thị 5 dòng dữ liệu đầu tiên (head) và cuối (tail) của ratings
   Cho biết kiểu dữ liệu (dtype) của các cột của ratings
   Xem thông tin (info) của ratings

Part 2: Xử lý dữ liệu bị thiếu/ không hợp lệ

1. Cho biết trong movies có dữ liệu null hay không? Nếu có loại bỏ dòng có dữ liệu null.
2. Cho biết trong ratings có dữ liệu null hay không? Nếu có loại bỏ dòng có dữ liệu null.
3. Cho biết trong tags có dữ liệu null hay không? Nếu có loại bỏ dòng có dữ liệu null.
4. Kiểm tra xem có dữ liệu rating nào không hợp lệ hay không ('rating' > 5 hoặc 'rating' < 0) ? Nếu có, hãy thay bằng giá trị xuất hiện nhiều nhất.

Part 3: Gộp DataFrame

1. Tạo movies_tags bằng cách gộp dữ liệu của movies và tags theo cột chung là ‘moviedId’. Hiển thị 5 dòng đầu của movies_tags.
2. Tạo movies_ratings bằng cách gộp dữ liệu của movies và ratings theo cột chung là ‘moviedId’. Hiển thị 5 dòng đầu của movies_ratings.

Part 4: Lọc dữ liệu theo yêu cầu

1. Tạo dataframe tag_counts cho biết với mỗi tag là có bao nhiêu film chứa giá trị ‘tag’ đó (gợi ý : dùng tags['tag'].value_counts()). Hiển thị 10 dòng đầu của tag_counts.
2. Tạo is_highly_rated theo điều kiện: có 'rating' >= 4.0 của dataframe ratings
   Hiển thị 5 dòng dữ liệu đầu của is_highly_rated. Liệt kê các phim thỏa is_highly_rated.
3. Tạo is_animation theo điều kiện trong cột genres của movies có chứa chuỗi 'Animation'
   Hiển thị 5 dòng dữ liệu đầu của is_animation. Liệt kê các phim thỏa is_animation.
4. Tạo movie_genres từ cột 'genres' bằng cách tách cột 'genres' dựa vào ký tự '|'
   Hiển thị 10 dòng cuối của movie_genres
5. Thêm cột mới cho movie_genres có tên là 'isComedy', giá trị là True nếu trong movies['genres'] có chứa chuỗi 'Comedy', ngược lại là False
   Hiển thị 10 dòng đầu của movie_genres
6. Thêm cột mới cho movies có tên là 'year' với year được lấy ra từ cột 'title' . Hiển thị 5 dòng dữ liệu đầu của movies

Part 5: Thống kê dữ liệu

1. Thực hiện thống kê chung dữ liệu ratings
2. In giá trị trung bình, giá trị lớn nhất, giá trị nhỏ nhất, độ lệch chuẩn, giá trị có tuần suất xuất hiện nhiều nhất của cột ‘rating’.
3. Thống kê đếm số lượng phim theo ‘rating’ (Count of films). Xem kết quả.
4. Đếm số lượng rating (Total ratings) theo phim, và lưu vào biến movie_count.
   Hiển thị 5 dòng dữ liệu đầu của movie_count
5. Tính rating trung bình (Average ratings) theo mỗi phim, và lưu vào biến avg_ratings.
6. Hiển thị rating trung bình của các phim là ‘Comedy’, chi in ra 5 dòng dữ liệu đầu của dataframe kết quả.

7. Hiển thị rating trung bình của các phim là ‘Comedy’ và có 'rating' >=4, chỉ in ra 5 dòng dữ liệu cuối của dataframe kết quả.
8. Tính trung bình rating theo year, và lưu vào biến yearly_average.
   Cho biết kích thước (shape) của yearly_average. Hiển thị 10 dòng dữ liệu đầu của yearly_average
9. Sắp xếp tăng dần theo cột year trong yearly_average. Hiển thị 20 dòng dữ liệu đầu của yearly_average_asc

Part 6: Parsing Timestamps

1. Tạo cột parsed_time trong tags bằng cách đổi thời gian sang định dạng DateTime
   Cho biết kiểu dữ liệu của cột parsed_time,
   Hiển thị 5 dòng dữ liệu của tags
2. Tạo selected_rows chứa các dòng có tags['parsed_time'] > '2015-02-01'.
3. Sắp xếp dữ liệu tags tăng dần theo cột parsed_time. Hiển thị 10 dòng dữ liệu đầu tiên của tags

Part 7: Trực quan hóa dữ liệu

1. Vẽ biểu đồ histogram cột 'rating' của ratings
2. Vẽ biểu đồ bar plot cột 'rating' của ratings
3. Dựa vào Câu 9 - Part 5, vẽ biểu đồ plot() cho 20 dòng đầu của yearly_average sau khi sắp tăng dần.
4. Quan sát biểu đồ trên và cho biết những năm nào có rating đặc biệt cao/thấp?