Đồ án 2 - Phân tích dữ liệu

Tổng quan:

Trang Soundcloud là một trang web cho phép người dùng upload và chia sẻ các bài hát. Chúng ta sẽ phân tích dữ liệu về các nghệ sĩ, ban nhạc, podcast và người sáng tác âm nhạc trên trang Soundcloud thông qua đồ án này.

Nhiệm vụ:

Thực hành các bước phân tích dữ liệu từ bộ dữ liệu Soundcloud thu thập được trong đồ án 01.

Mục tiêu:

Làm quen và biết cách xử lý dữ liệu cơ bản, phân tích và rút ra các insight về tập dữ liệu.

Chi tiết

Trước khi khám phá dữ liệu, hãy viết ra danh sách ngắn về những gì bạn mong đợi sẽ thấy trong dữ liệu: sự phân bố của các biến chính, mối quan hệ / tương quan giữa các cặp biến, v.v. Danh sách này về cơ bản là một dự đoán dựa trên hiểu biết hiện tại của bạn về dữ liệu.

Trong bước phân tích dữ liệu, lập bảng, tóm tắt, bất cứ điều gì cần thiết để xem nó có phù hợp với mong đợi của bạn không.

- Danh sách kiểm tra phân tích dữ liệu: Danh sách kiểm tra này có thể được sử dụng như một hướng dẫn trong quá trình phân tích dữ liệu hoặc như một cách để đánh giá chất lượng của một phân tích dữ liệu được báo cáo.

- Trả lời những câu hỏi về bộ dữ liệu:
    1. Bạn đã xác định số liệu trước khi bắt đầu?
    2. Bạn đã hiểu ngữ cảnh cho câu hỏi và ứng dụng?
    3. Bạn đã xem xét liệu câu hỏi có thể được trả lời với dữ liệu có sẵn không?

- Xóa dữ liệu:

1. Dữ liệu có bị thiếu không?
2. Mỗi bảng có các kiểu dữ liệu khác nhau? Có kiểu dữ liệu nào chưa phù hợp?
3. Kiểm tra các ngoại lệ

- Phân tích khám phá:
    1. Trực quan hoá mối quan hệ đơn biến (histogram, distplot, boxplot)
    2. Trực quan hoá các tương quan đa biến (scatterplot, jointplot, kde plot, correlation matrix)

- Trình bày:
    1. Bạn đã dẫn dắt một cách ngắn gọn, dễ hiểu cho mọi người về vấn đề của bạn?
    2. Bạn đã giải thích dữ liệu, mô tả câu hỏi cần quan tâm?

Yêu cầu

1. Code

Làm trực tiếp trên các file notebook .ipynb. Các bạn có thể sử dụng jupyter lab trong quá trình thực hiện nếu thấy thuận tiện.

2. Báo cáo

Viết trực tiếp trong các ô markdown của file notebook đã code.

3. Bảng phân công công việc

Đầu notebook của mỗi nhóm cần có bảng danh sách tên và phân công công việc của các thành viên.
Các bạn lưu ý chia việc hợp lý, mỗi thành viên cần đảm nhận lượng công việc tương đương nhau.

4. Lưu ý

Bài làm giống nhau 0 điểm cả môn.
Ghi rõ nguồn tham khảo đầy đủ.
Cần trình bày phần báo cáo rõ ràng, dễ hiểu.

5. Nộp bài

Folder bài nộp cần có:

- File notebook của phần code (.ipynb)
- Bản pdf của các file notebook trên (export file .ipynb thành PDF)

Nén folder thành một file, đặt tên theo cú pháp sau và nộp qua moodle:

<MSSV1>_<MSSV2>_<MSSV3>_<MSSV4>_<MSSV5>.zip

Thông tin liên hệ

TA: Nguyễn Ngọc Băng Tâm

Nếu có thắc mắc các bạn vui lòng liên hệ qua: bangtamnguyenn@gmail.com