TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

ĐẠI HỌC QUỐC GIA – THÀNH PHỐ HỒ CHÍ MINH

-----o0o----

NHẬP MÔN KHOA HỌC DỮ LIỆU

REPORT

FINAL PROJECT

Nhóm sinh viên thực hiện:

20120041 – Trần Kim Bảo

20120053 – Nguyễn Thành Đạt

20120071 – Nguyễn Thị Bích Hà

20120113 – Lê Nguyên Khang

Thành phố Hồ Chí Minh, tháng 12 năm 2022

MỤC LỤC

I. BẢNG THÀNH VIÊN NHÓM

MSSV	Họ và tên
20120041	Trần Kim Bảo
20120053	Nguyễn Thành Đạt
20120071	Nguyễn Thị Bích Hà
20120113	Lê Nguyên Khang

II. BẢNG PHÂN CÔNG CÔNG VIỆC

Về phần cào dữ liệu và soạn báo cáo, tất cả các thành viên đều làm chung.

MSSV	Nội dung	Mức độ hoàn thành
20120041	• Mỗi dòng có ý nghĩa gì? Có vấn đề các dòng có ý nghĩa khác nhau không? • Đặt và trả lời câu hỏi 1. • Làm và nguyên cứu phần mô hình (Mô hình phân cụm).	100%
20120053	• Mỗi cột hiện đang có kiểu dữ liệu gì? Khoảng biểu diễn của kiểu dữ liệu đó ra sao? Có cột nào có kiểu dữ liệu chưa phù hợp để có thể xử lý tiếp hay không? • Đặt và trả lời câu hỏi 3. • Làm và nguyên cứu phần mô hình (Mô hình phân cụm).	100%

20120071	• Mỗi cột có ý nghĩa gì? • Đặt và trả lời câu hỏi 2, 4. • Tổng hợp và đồng bộ hóa báo cáo (Slide và PDF). • Merge tay file jupyter và kiểm tra lỗi trình bày, chính tả.	100%
20120113	• Với mỗi cột, dữ liệu phân bố như thế nào? • Đặt và trả lời câu hỏi 5. • Làm và nghiên cứu phần mô hình (Mô hình hồi quy tuyến tính).	100%

III. ĐÁNH GIÁ CÔNG VIỆC

1. 20120041 – Trần Kim Bảo

A. Khó khăn

- Đặt câu hỏi: Tìm kiếm ý tưởng khá lâu
- Trả lời câu hỏi: Đưa ra hướng giải quyết, suy nghĩ liệu hướng giải quyết của bản thân có đủ lập luận để trả lời câu hỏi tổng thể hay không.
- Github: Với lần đầu sử dụng, nên gặp phải những trường hợp:
  - Pull nhưng code không về
  - Commit trực tiếp trên terminal của ubuntu không được (sau em giải quyết bằng cách commit trong vscode).
  - Merge code tay nên, bị thiếu cell dữ liệu.
- Lấy dữ liệu từ Website:

- Có ô chưa dữ liệu bị trống, nên có dữ liệu bị đưa vào sai cột
- Bố cục của trang web bị thay đổi, lấy về bị null.

- Mô hình hóa dữ liệu :
  - Khó trong định hướng làm bài, nên sử dụng thuật toán nào.
  - Chưa hiểu rõ cách hoạt động của mô hình, dẫn đến khó tìm và đọc tài liệu tham khảo.

B. Điều học được

- Biết cách sử dụng Git (còn quản lý code trên này thì em chưa nắm được)
- Trong qua trình xử lý dữ liệu: hiểu rõ hơn về thư viện regex, có sự thay đổi trong các dòng code (clean hơn lúc trước, và dễ hiểu hơn).
- Học hỏi được cách xử lý thông tin, hướng giải quyết từ các bạn trong nhóm.

C. Nếu có thời gian sẽ làm

- Sẽ tìm hiểu thêm về cách đọc biểu đồ, và đưa ra nhận xét
- Tìm hướng tối ưu xử lý phần dữ liệu chưa hoàn thiện.

2. 20120053 – Nguyễn Thành Đạt

A. Khó khăn

- Em gặp khó khăn trong việc tìm câu hỏi và tìm hướng giải quyết câu hỏi làm sao cho thuyết phục.
- Em chưa sắp xếp thời gian ổn thỏa được giữa các môn với nhau nên ảnh hưởng khá nhiều đến quá trình thực hiện.

B. Điều học được

- Em học được thêm một số cách trực quan hóa và nhận xét từ biểu đồ đã trực quan.
- Ngoài ra em còn biết được thêm một số thao tác sử dụng Git đơn giản vì trước giờ em rất ít khi sử dụng công cụ này.
- Em biết được thêm một số cách xử lý dữ liệu hay mà lúc trước em chưa từng biết.

C. Nếu có thời gian sẽ làm

- Đọc kĩ dữ liệu hơn, cào thêm một số trường thông tin nữa để đa dạng hóa việc phân tích.
- Tìm hiểu về Git để thao tác mượt mà hơn hiện tại.
- Xem thêm một số biểu đồ để có thể trực quan hóa theo nhiều hướng đa dạng hơn hiện tại.

3. 20120071 – Nguyễn Thị Bích Hà

A. Khó khăn

- Em gặp khó khăn trong việc đề ra câu hỏi vì phân vân không biết mình sẽ xử lý câu hỏi đó như thế, lập luận ra sao để mang tính thuyết phục. Nên đã tốn rất nhiều thời gian cho phần này và làm trễ những phần khác.
- Vì lần đầu dùng Git nên em không biết phải làm sao để up file lên, chia brand, merge. Và đã phải merge tay hết các cell với các bạn. Sau đó em lỡ up hết 1 file jupyter gần hoàn thành lên, em đang phân vân không biết điều này có ảnh hưởng đến các bạn khác không.
- Em còn gặp một số khó khăn trong việc xử lý dữ liệu, tiền xử lý, tính toán dữ liệu (thiên về kĩ thuật).

B. Điều học được

- Em đã thành thạo hơn một số kĩ thuật clean data cũng như xử lý dữ liệu khi phân tích và trả lời câu hỏi.
- Biết thêm một số thao tác đơn giản để sử dụng Git.
- Biết thêm một số biểu đồ khác phục vụ cho việc visualize phân tích.
- Biết thêm được một số thông tin bổ ích từ data trong lúc thực hiện đặt và trả lời câu hỏi và luyện tập thêm được khả năng nhìn biểu đồ và phân tích.

C. Nếu có thời gian sẽ làm

- Em sẽ tìm hiểu kĩ hơn về dataset đang sử dụng và đặt ra một câu hỏi đòi hỏi xử lý và phân tích nhiều hơn câu hỏi mà em đã làm.
- Em sẽ tìm hiểu thêm một số dạng biểu đồ mới để phục vụ cho việc phân tích dữ liệu.
- Học cách sử dụng Git để có thể sử dụng thành thạo hơn và tránh gây ra lỗi cho các thành viên khác.
- Xem thêm phần mô hình để hoàn thiện bài làm hơn.

4. 20120113 – Lê Nguyên Khang

A. Khó khăn

- Đơn vị tiền tệ trong bài được sử dụng khá nhiều loại khó khăn trong quá trình chuyển đổi xử lý khi tiền xử lý dữ liệu.
- Khó khăn tìm hướng giải quyết câu hỏi sao cho rõ ràng và thuyết phục
- Việc chia phần quá nhỏ khiến việc merge code trở nên khó khăn và dễ sai sót.
- Không đồng nhất các ý tưởng giữ thành viên (nhóm quyết định làm 2 bài toán khác nhau)

B. Điều học được

- Em học được thêm một số cách clean dữ liệu.

- Nhờ vào việc phân tích các câu hỏi, em biết thêm được một số thông tin bổ ích về dữ liệu hiện tại.
- Biết sử dụng các dạng biểu đồ khác nhau để nhận xét tương quan giữa các cột dữ liệu.
- Nhạy bén trong việc tư duy phát hiện một yếu tố trong cuộc sống sẽ phụ thuộc vào yếu tố nào khác trong cuộc sống.
- Học cái mới bằng cách tự học không phải là không thể, chỉ cần tìm tòi chăm chỉ không ngừng là sẽ có thể làm được.

C. Nếu có thời gian sẽ làm

- Cào thêm một số trường dữ liệu để có thêm nhiều thông tin trong việc đặt câu hỏi hơn.
- Tìm hiểu nhiều hơn về các loại mô hình khác nhau nhằm phụ hợp, hợp lý hơn với bài toán đặt ra.