Đồ án cuối kỳ
-Xử lý số liệu thống kê-

TS.Tô Đức Khánh

15/12/2024

Chọn 1 trong các project dưới đây. Chú ý, với bất kỳ project nào được chọn, nhóm cần phải hoàn thành các yêu cầu sau:

1. Bản đề xuất phân tích và xử lý số liệu dựa trên các phương pháp đã được học trong học phần.
2. Các mục tiêu phân tích cần đạt được.
3. Các phương pháp và chiến lược (các bước) phân tích cho mỗi mục tiêu đã đề ra.
4. Mô tả và biểu diễn tổng hợp dữ liệu (bảng tổng hợp, biểu đồ).
5. Phân tích để đạt được các mục tiêu đã đề ra, kết quả được biểu diễn dưới dạng bảng tổng hợp, biểu đồ.
6. Viết các nhận xét và kết luận về các kết quả đã thu được sau quá trình phân tích.

Toàn bộ quá trình được tổng hợp lại dưới dạng file báo cáo (.pdf, không kèm code), code R và kết quả được nộp ở file Rmd và html

1 Project 1 - Sports Data Analysis

Mục tiêu của dự án này là phân tích đánh giá cầu thủ đựa trên các thông tin về tiền lương, quốc tịch, độ tuổi, câu lạc bộ mà họ hiện đang chơi và nhiều biện pháp đánh giá hiệu suất khác nhau. Việc phân tích đánh giá này sẽ giúp cho ban quản lý của câu lạc bộ đưa ra các quyết định mua sắm cầu thủ hợp lý dựa trên ngân sách của câu lạc bộ.

Dữ liệu gồm thông tin của 18207 cầu thủ, được tổng hợp trong 01 file dữ liệu `fifa_eda_stats.csv`, bao gồm 57 biến, chẳng hạn:

- `ID` - mã số của cầu thủ;
- `Name` - tên cầu thủ;
- `Age` - tuổi;
- `Nationality` - quốc tịch;
- `Overall` - điểm đánh giá tổng thể (tối đa 100);
- `Potential` - điểm đánh giá tiềm năng (tối đa 100);
- `Club` - tên câu lạc bộ đang chơi;
- `Value` - giá trị trên thị trường chuyển nhượng;
- `Wage` - tiền lương;
- `Preferred.Foot` - chân thuận;
- `Release.Clause` - chi phí giải phóng hợp đồng;

- `Height` - chiều cao;
- `Weight` - cân nặng;
- `Position` - vị trí thi đấu sở trường;
- và các biến khác đo các chỉ số đánh giá.

Chú ý có một số biến bị sai định dạng khi nhập vào file lưu trữ (số nhưng lưu ở dạng chữ), do đó cần hiệu chỉnh lại cho đúng trước khi xử lý chính.

2 Project 2 - CDC Diabetes Health Indicators

Hiện nay, bệnh tiểu đường là một căn bệnh mãn tính phổ biến nhất trên thế giới. Bệnh tiểu đường là một căn bệnh mãn tính nghiêm trọng khiến mọi người mất khả năng điều chỉnh hiệu quả lượng glucose trong máu và có thể dẫn đến giảm chất lượng cuộc sống và tuổi thọ. Các biến chứng như bệnh tim, mất thị lực, cắt cụt chi dưới và bệnh thận có liên quan đến lượng đường cao mãn tính vẫn còn trong máu đối với những người mắc bệnh tiểu đường. Mặc dù không có cách chữa khỏi bệnh tiểu đường, nhưng các chiến lược như giảm cân, ăn uống lành mạnh, vận động và điều trị y tế có thể làm giảm tác hại của căn bệnh này ở nhiều bệnh nhân. Chẩn đoán sớm có thể dẫn đến thay đổi lối sống và điều trị hiệu quả hơn, khiến các mô hình dự đoán nguy cơ mắc bệnh tiểu đường trở thành công cụ quan trọng đối với cộng đồng và các quan chức y tế công cộng.

Dữ liệu `diabetes_012_health_indicators_BRFSS2015.csv` chứa thông tin khảo sát của 253,680 người dân Hoa Kỳ (năm 2015), với 22 biến được quan sát, chẳng hạn:

- `Diabetes_012` - tình trạng bệnh tiểu đường (0: không tiểu đường, 1: tiền tiểu đường, 2: tiểu đường);
- `HighBP` - tình trạng cao huyết áp (0/1);
- `HighChol` - tình trạng cao cholesterol (0/1);
- `CholCheck` - kiểm tra cholesterol trong 5 năm (0/1);
- `BMI` - Body Mass Index
- `Smoker` - người đã hút ít nhất 100 điếu thuốc trong suốt cuộc đời mình (0/1), [lưu ý: 5 gói = 100 điếu thuốc];
- `Stroker` - đã từng đột quỵ (0/1);
- `HeartDiseaseorAttack` - bệnh tim mạch vành (CHD) hoặc nhồi máu cơ tim (MI), 0/1;
- `PhysActivity` - hoạt động thể chất trong vòng 30 ngày, không tính hoạt động liên quan tới công việc, 0/1;
- `Sex` - giới tính (0: nữ, 1: nam);
- `Age` - tuổi (13 nhóm, xem thêm trong danh mục tra cứu);
- `Education` - cấp giáo dục đã hoàn thành (xem thêm trong danh mục tra cứu);
- `Income` - mức thu nhập (xem thêm trong danh mục tra cứu);
- `GenHlth` - điểm tự đánh giá sức khỏe chung (xem thêm trong danh mục tra cứu);
- và các biến sức khỏe khác, xem chi tiết tại https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?select=diabetes_012_health_indicators_BRFSS2015.csv

Danh mục tra cứu biến tại https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf

Chú ý tới số lượng người trong mỗi nhóm tình trạng bệnh tiểu đường.

3 Project 3 - Body performance Data

Hiện nay các phong trào tập thể thao đang ngày một phát triển, thu hút nhiều nhóm tuổi và giới tính. Dữ liệu `bodyPerformance.csv` chứa thông tin của 13,393 người tham gia tập thể thao tại Hàn Quốc, với 12 biến như sau:

- `age` - độ tuổi (từ 20 tới 64);
- `gender` - giới tính (F: nữ, M: nam);
- `height_cm` - chiều cao (đơn vị: cm);
- `weight_kg` - cân nặng (đơn vị: kg);
- `body fat_%` - phần trăm mỡ cơ thể (%);
- `diastolic` - huyết áp tâm trương (phút);
- `systolic` - huyết áp tâm thu (phút);
- `gripForce` - lực kẹp;
- `sit and bend forward_cm` - ngồi và gập người về phía trước;
- `sit-ups counts` - số lần gập bụng;
- `broad jump_cm` - nhảy xa (đơn vị: cm);
- `class` - phân lớp hiệu suất (A: tốt nhất, B,C,D).

Hãy xử lý dữ liệu này để giúp cho các chuyên gia sức khỏe biết được hiệu quả của việc tập thể dục, và các yếu tố ảnh hưởng tới hiệu quả.

4 Project 4 - Turkish Crowdfunding Startups

Bộ dữ liệu `turkishCF.xlxs` chứa dữ liệu về các chiến dịch gây quỹ cộng đồng (Crowdfunding) tại Thổ Nhĩ Kỳ. Bộ dữ liệu bao gồm 38 biến miêu tả đặc điểm khác nhau như các dự án gây quỹ cộng đồng, mô tả dự án, quỹ mục tiêu và quỹ đã huy động được, thời lượng chiến dịch và số lượng người ủng hộ. Được thu thập vào năm 2022, bộ dữ liệu này cung cấp một nguồn tài nguyên có giá trị cho các nhà nghiên cứu muốn hiểu và phân tích hệ sinh thái gây quỹ cộng đồng tại Thổ Nhĩ Kỳ. Tổng cộng, có dữ liệu từ hơn 1,500 dự án trên 6 nền tảng khác nhau. Bộ dữ liệu này là điểm tham chiếu quan trọng cho các nghiên cứu về đặc điểm của các chiến dịch gây quỹ cộng đồng thành công và cung cấp thông tin toàn diện cho các doanh nhân, nhà đầu tư và nhà nghiên cứu tại Thổ Nhĩ Kỳ. Chi tiết các biến xem tại file `dictionary_turkishCF.txt`.

Chú ý, một số biến định tính được lưu trữ bằng tiếng Thổ Nhĩ Kỳ, do đó cần 1 bước chuyển đổi (nếu cần). Một số biến quan trọng:

- `basari_durumu` - gây quỹ thành công: başarılı (thành công), başarısız (không thành công);
- `toplanan_tutar` - số tiền tài trợ thu được cho dự án;
- `hedef_miktari` - mục tiêu số tiền tài trợ cho dự án;
- `proje_sahibi_cinsiyet` - giới tính của chủ dự án: belirsiz (không rõ), erkek (nam), kadın (nữ);
- `video_uzunlugu` - chiều dài của video quảng cáo gây quỹ;
- `icerik_kelime_sayisi` - số lượng từ trong mô tả dự án.