CÁC PHƯƠNG PHÁP ĐỊNH LƯỢNG 1

BÀI THI GIỮA KỲ

Ngày phát: 02/12/2022

Hạn nộp: 8h20, 04/12/2022

Bài làm được yêu cầu chỉ nộp bản điện tử trên Microsoft Teams

---

File ***demoSubsetHTS.csv*** chứa dữ liệu về đặc tính nhân khẩu học của 20.000 người (được rút ngẫu nhiên từ tổng số hơn 60.000 người) tham gia điều tra về di chuyển hộ gia đình tại Thành phố Hồ Chi Minh, do Cơ quan Hợp tác Quốc tế Nhật Bản (JICA) tài trợ thực hiện vào 2013 - 2014.

Các cột của bộ dữ liệu này được mô tả ngắn gọn dưới đây:

- id - mã định danh của người tham gia khảo sát;
- age - tuổi của người tham gia khảo sát;
- gender - giới tính của người tham gia khảo sát;
- income - thu nhập mỗi tháng của người tham gia khảo sát;
- eduLevel - trình độ học vấn cao nhất đạt được của người tham gia khảo sát, có thể có các giá trị dưới đây:
  - juniorHighOrLower: Tốt nghiệp trung học cơ sở hoặc thấp hơn;
  - seniorHighAndEquivalent: Tốt nghiệp trung học phổ thông hoặc tương đương;
  - tertiaryEduOrHigher: Tốt nghiệp đại học, cao đẳng hoặc cao hơn;
- licence - có bằng lái xe cơ giới, có thể có các giá trị dưới đây:
  - yes: có ít nhất 1 bằng lái xe cơ giới;
  - no: không có bằng lái xe cơ giới;
- jobType - loại công việc của người tham gia khảo sát, có thể có các giá trị dưới đây:
  - officeWorker: làm việc văn phòng;
  - labourWorker: công nhân hoặc người làm việc thủ công;
  - student: sinh viên học sinh;
  - unemployed/Retired: không có việc làm hoặc đã nghỉ hưu;
  - other: các loại hình công việc khác;
- employStatus - tình trạng việc làm của người tham gia khảo sát, có thể có các giá trị dưới đây:
  - short-termOrFreelancer: công việc ngắn hạn hoặc việc làm tự do;
  - permanentFulltime: công việc toàn thời gian;
  - businessOwner: chủ doanh nghiệp;
- ownMotorVehicles - có sở hữu xe máy hoặc ôtô, có thể có các giá trị dưới đây:
  - yes: sở hữu tối thiểu 1 xe máy hoặc 1 ôtô;
  - no: không sở hữu bất kỳ xe máy hoặc ôtô;

Anh/chị phân tích khám phá bộ dữ liệu này và trình bày các đặc điểm của bộ dữ liệu trong giới hạn khoảng 3.000 từ (dao động +/-5%) dựa trên các yêu cầu cụ thể dưới đây:

1. Anh/Chị trình bày đặc điểm của từng thuộc tính nhân khẩu học dựa trên các trị thống kê mô tả và các đồ thị tương ứng.

2. Anh/Chị xác định và phân loại các giá trị bất thường có thể có (outliers, inliers, missing values) trong từng thuộc tính nhân khẩu học. Anh/Chị giải thích đầy đủ các bước và các tính toán (nếu có) trong việc xác định các giá trị bất thường này.

3. Anh/Chị thực hiện thay thế các giá trị được xem là bị mất trong các cột giá trị trên. Giải thích đầy đủ các bước và các tính toán (nếu có) của việc xác định các giá trị bị mất. Anh/chị có nhận định gì về đặc tính của các thuộc tính nhân khẩu học trước và sau khi thay thế giá trị bị mất? Minh họa các nhận định này của anh/chị bằng các trị thống kê, bảng biểu, hoặc đồ thị phù hợp.

4. Anh/chị vẽ các đồ thị phù hợp để mô tả sự tương quan giữa các cặp biến sau:

- income và age
- income và jobType
- income và licence

5. Theo anh/chị, dữ liệu khảo sát có cung cấp đủ bằng chứng cho thấy rằng có sự khác nhau giữa thu nhập trung bình (số liệu trong cột income) của người có bằng lái xe và không có bằng lái xe không (số liệu trong cột licence)? Anh/Chị trình bày chi tiết các bước của bài kiểm định thống kê phù hợp để giải thích cho kết luận của mình.

6. Anh/chị kiểm định giả thuyết cho rằng có sự tương quan giữa 2 biến eduLevel và jobType (sử dụng số liệu sau khi đã thay thế giá trị bị mất, nếu có). Anh/Chị trình bày chi tiết các bước của bài kiểm định, tự chọn mức kiểm định và đưa ra kết luận tương ứng.

- **Lưu ý:**

- Anh/Chị chỉ cần thực hiện thay thế dữ liệu bị mất 1 lần (single imputation) và với giả sử rằng việc mất dữ liệu là hoàn toàn ngẫu nhiên (missing completely at random).

- Anh/Chị có thể dùng công cụ tính toán phù hợp. Tuy nhiên, anh/chị cần trình bày các bước phân tích và kết quả đầy đủ và súc tích, hạn chế đưa các hàm hoặc các dòng code trực tiếp vào báo cáo. Thay vào đó, anh/chị mô tả chức năng, thuộc tính, các giả định, và các tham số đầu vào và các kết quả của các hàm hoặc các dòng code đó.

- Anh/Chị cần nộp báo cáo và đầy đủ các file tính toán (tức file excel hoặc codes). Tuy nhiên, anh/chị lưu ý rằng việc chấm điểm hoàn toàn chỉ dựa trên các thông tin được trình bày trong báo cáo. Các file tính toán chỉ được dùng để tham khảo và kiểm chứng các kết quả được trình bày trong báo cáo trong trường hợp cần thiết.

---HẾT---