**PHÂN TÍCH DỮ LIỆU**

**Bài 3**

Bộ dữ liệu (Visastatus.csv) mô tả thông tin về những người nộp hồ sơ xin visa ở Mỹ. Các biến của bộ dữ liệu được mô tả như bảng 1.

Bảng 1: Mô tả dữ liệu

|  |  |
| --- | --- |
| **Tên cột (biến)** | **Giải thích** |
| Order | Số thứ tự |
| CASE\_NUMBER | Mã hồ sơ |
| CASE\_STATUS | Tình trạng hồ sơ |
| CASE\_RECEIVED\_DATE | Ngày nhận hồ sơ |
| DECISION\_DATE | Ngày có quyết định |
| EMPLOYER\_NAME | Tên công ty |
| JOB\_TITLE | Nghề nghiệp |
| WORK\_CITY | Thành phố làm việc |
| EDUCATION\_LEVEL\_REQUIRED | Trình độ học vấn |
| COLLEGE\_MAJOR\_REQUIRED | Ngành học |
| COUNTRY OF CITIZENSHIP | Quốc tịch |
| WORK\_STATE | Bang làm việc |
| WORK\_STATE\_ABBREVIATION | Kí hiệu bang làm việc |
| VISA\_CLASS | Visa xin được cấp |
| PREVAILING\_WAGE\_PER\_YEAR | Mức lương kì vọng |
| PAGE\_WAGE\_PER\_YEAR | Mức lương thực tế |
| JOB\_TITLE\_SUBGROUP | Nghề nghiệp theo nhóm |

Yêu cầu:

1. Sử dụng thư viện Table1, lập bảng tần số phân phối số lượng người nộp hồ sơ theo bang của từng nhóm nghề. Bang nào có nhiều người nộp hồ sơ nhất? Bang nào có ít người nộp hồ sơ nhất?
2. Có bao nhiêu loại visa được cấp (cột VISA\_CLASS)? Loại visa nào có nhiều người nộp nhất? Loại visa nào có ít người nộp nhất?
3. Lập bảng tần số phân phối số lượng người nộp hồ sơ theo từng nhóm nghề.
4. Vẽ biểu đồ hình tròn mô tả tỉ lệ phần trăm số lượng người nộp hồ sơ theo từng nhóm nghề.
5. Vẽ biểu đồ hình cột mô tả số lượng người nộp hồ sơ theo từng nhóm nghề.
6. Trong ba công cụ trực quan hóa dữ liệu (bảng, biểu đồ hình cột, biểu đồ hình tròn), công cụ nào phản ánh thông tin hiệu quả nhất? Vì sao?
7. Vẽ biểu đồ tần số phân phối (histogram) mô tả về phân phối mức lương thực tế trong mẫu nghiên cứu và mức lương kì vọng trong từng mẫu nghiên cứu. Hãy cho biết hình dạng của phân phối và cho biết bạn lựa chọn tham số trung tâm nào để phản ánh vị trí trung tâm của mức lương thực tế và mức lương kỳ vọng. Vì sao? Phân phối của mức lương thực tế và mức lương kỳ vọng có giá trị ngoại lai không? Là những giá trị nào?
8. Vẽ biểu đồ hộp (boxplot) mô tả mức lương thực tế và mức lương kỳ vọng. Nhận xét về hình dạng của các phân phối. Nhận diện giá trị ngoại lai và so sánh với kết quả ở câu 6.
9. Vẽ biểu đồ tần suất Kernel cho mức lương thực tế và cho biết có bao nhiêu phần trăm người nộp hồ sơ có mức lương thực tế tháp hơn mức trung bình của toàn bộ mẫu.