

THỐNG KÊ MÁY TÍNH & ỨNG DỤNG – 18HCB
BÀI TẬP 3
---oOo---

Phân tích dữ liệu thăm dò (Exploratory data analysis) hay *thống kê mô tả* (Descriptive statistics) là bước phân tích đầu tiên (sau khi thu thập, xử lý dữ liệu) để “cảm nhận”, “nhìn thấy” dữ liệu trước khi thực hiện các phân tích sâu hơn (định hướng hơn, và tốn kém hơn) trên dữ liệu. Công cụ chủ yếu để thực hiện phân tích dữ liệu thăm dò là các *thống kê tóm tắt* (summary statistic) và *biểu đồ trực quan* (chart/graph/plot).

Hướng dẫn







Sinh viên thực hành theo tài liệu: Chapter 2 – Describing Data, Randall Pruim and Lana Park, *Lock5 with R: A companion to Unlocking the Power of Data*, 2014.

Sinh viên đọc thêm lý thuyết trong tài liệu: Chapter 2 – Describing Data, Robin H. Lock et al. *Statistics: Unlocking the power of data*, Wiley, 2013.

Bài tập nộp

Sinh viên trình bày bài trong file Word (bao gồm code R, kết quả phân tích, biểu đồ, ...) và nộp bài vào link nộp bài trên trang môn học.

Sinh viên thực hiện các yêu cầu sau trên dataset **SpeedDating** trong package **Lock5withR**.

- 1) Mô tả tổng quan dataset (nguồn gốc, mục đích nghiên cứu, tổng thể nghiên cứu, cách thu thập dữ liệu, cỡ mẫu, số lượng biến).
-  2) Chọn ra 5 biến quan tâm, trong đó có ít nhất 2 biến định tính và 2 biến định lượng. Mô tả sơ lược ý nghĩa 5 biến này và nêu lí do chọn.
-  3) Phân tích thăm dò riêng từng biến đã chọn. 
-   4) Chọn ra 2 biến định tính (từ 5 biến quan tâm) và phân tích thăm dò quan hệ giữa chúng.
-  5) Chọn ra 1 biến định tính, 1 biến định lượng (từ 5 biến quan tâm) và phân tích thăm dò quan hệ giữa chúng.
- 6) Chọn ra 2 biến định lượng (từ 5 biến quan tâm) và phân tích thăm dò quan hệ giữa chúng.
- 7) (*Cộng điểm*) Phân tích thăm dò quan hệ giữa nhiều hơn 2 biến (từ 5 biến quan tâm).

Việc phân tích thăm dò bao gồm:

- tóm tắt bằng các thống kê phù hợp
- trình bày trực quan bằng các biểu đồ phù hợp
- đưa ra các nhận xét, đánh giá, nhận định ban đầu (“cảm tính”) trên các biến hay quan hệ giữa các biến
- có thể nêu thêm các mở rộng, nghi vấn quan tâm để tiếp tục phân tích sâu hơn sau này