

Các xử lý cơ bản (1)

Trịnh Tấn Đạt

Đại Học Sài Gòn

trinhtandat@sgu.edu.vn

<http://sites.google.com/site/ttdat88>

Nội Dung

- Phân tích dữ liệu

Phân tích dữ liệu

- Ứng dụng pandas trong phân tích dữ liệu
- Đọc dữ liệu về giá nhà ở tại Boston từ bộ dữ liệu BostonHousing.
- Tham khảo

<https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset/notebook>

<https://github.com/selva86/datasets/blob/master/BostonHousing.csv>

• The following describes the dataset columns:

- CRIM - per capita crime rate by town
- ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS - proportion of non-retail business acres per town.
- CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- NOX - nitric oxides concentration (parts per 10 million)
- RM - average number of rooms per dwelling
- AGE - proportion of owner-occupied units built prior to 1940
- DIS - weighted distances to five Boston employment centres
- RAD - index of accessibility to radial highways
- TAX - full-value property-tax rate per \$10,000
- PTRATIO - pupil-teacher ratio by town
- B - $1000(B_k - 0.63)^2$ where B_k is the proportion of blacks by town
- LSTAT - % lower status of the population
- MEDV - Median value of owner-occupied homes in \$1000's
- crim: Tỷ lệ phạm tội phạm bình quân đầu người theo thị trấn.
- zn: Tỷ lệ đất ở được quy hoạch cho các lô trên 25.000 foot square.
- indus: Tỷ lệ diện tích thuộc lĩnh vực *kinh doanh phi bán lẻ* trên mỗi thị trấn.
- chas: Biến giả, = 1 nếu được bao bởi sông Charles River, = 0 nếu ngược lại.
- nox: Nồng độ khí Ni-tơ oxit.
- rm: Trung bình số phòng trên một căn hộ.
- age: Tỷ lệ căn hộ được xây dựng trước năm 1940.
- dis: Khoảng cách trung bình có trọng số tới 5 trung tâm việc làm lớn nhất ở Boston.
- rad: Chỉ số về khả năng tiếp cận đường cao tốc.
- tax: Giá trị thuế suất tính trên đơn vị **10000\$**.
- ptratio: Tỷ lệ học sinh-giáo viên trên mỗi thị trấn.
- black: Tỷ lệ số người da đen trong thị trấn được tính theo công thức: $1000(B_k - 0.63)^2$ ở đây B_k là tỷ lệ người da đen trong thị trấn.
- lstat: Tỷ lệ phần trăm dân số thu nhập thấp.
- medv: median giá trị của nhà có người sở hữu tính trên đơn vị **1000\$**.

Tham khảo

- <https://drive.google.com/file/d/1qJtcOV-DgGxo8zsJnLgSAydzgqlxaibv/view?usp=sharing>
- <https://drive.google.com/file/d/1RRBFtHiRhVS8Z2kqc-vA7UG2IpAomCRb/view?usp=sharing>