# DATA UNDERSTANDING AND VISUALIZATION

Nguyễn Thị Minh Tâm - 18021113

Nhóm 5

K63 - CACLC3

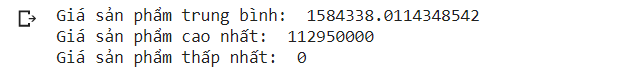
### Tổng quan dữ liệu

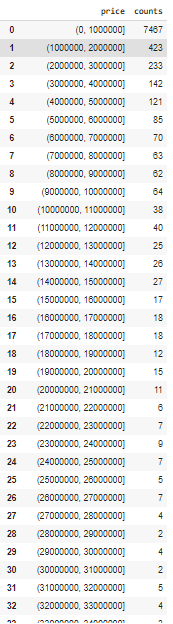
* File csv: <https://github.com/tamtamxtamtam/NguyenThiMinhTam_18021113_Nhom5_Crawler>
* Mô tả: Dữ liệu crawl được là các sản phẩm của trang TMĐT Tiki.vn, bao gồm khoảng 8000 sản phẩm trong 16 ngành hàng khác nhau. Dữ liệu đã được làm sạch (chuyển các đoạn text chứa giá tiền, số lượng về số thuần,...). Định dạng “.csv”.
* Chi tiết file:
* name: Tên sản phẩm
* sku: Mã sản phẩm
* no\_of\_stars: số đánh giá sao trung bình của sản phẩm ( từ 0-5 sao; tính bằng tổng số sao của mỗi đánh giá hợp lệ chia cho số lượng đánh giá)
* no\_of\_comments: Số lượt đánh giá
* seller\_name: Tên nhà cung cấp
* seller\_link: Link của nhà cung cấp
* img\_link: Ảnh của sản phẩm
* price: Giá của sản phẩm sau khi đã giảm giá (nếu có)
* original\_price: Giá gốc của sản phẩm (hoặc giá thị trường)
* brief: Giới thiệu tóm tắt
* description: Thông số chi tiết sản phẩm (nếu có)
* info: Giới thiệu chi tiết sản phẩm

### Thư viện được sử dụng

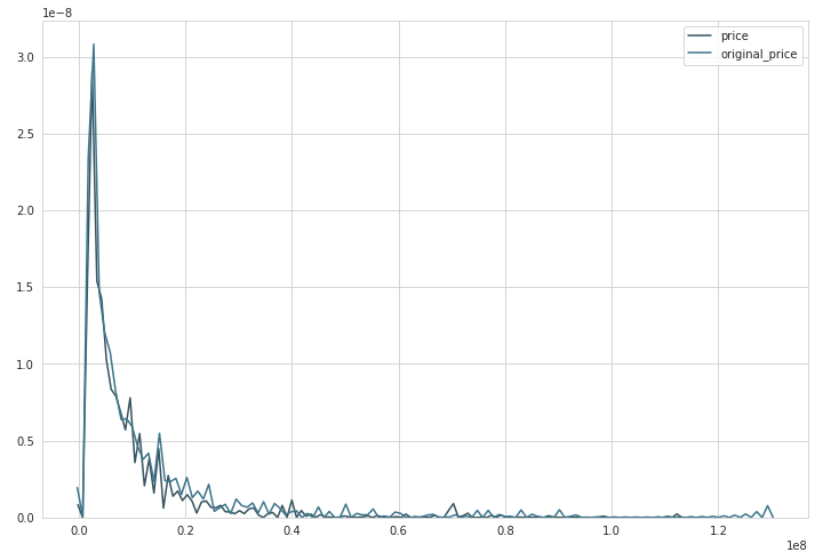
* Pandas
* Seaborn

### Đánh giá khoảng giá tiền của sản phẩm



* Các sản phẩm có khoảng cách giá lớn, xét theo giá tiền trung bình có thể dự đoán các sản phẩm có giá trị nhỏ chiếm phần nhiều trong tổng dữ liệu.
* Chia nhỏ giá sản phẩm thành các khoảng cách nhau 1.000.000 vnđ; lọc bỏ các khoảng giá không có sản phẩm nào. Ta có thể thấy số lượng sản phẩm đúng như dự đoán, các sản phẩm chủ yếu tập trung giá trong khoảng từ 0 - 15.000.000 vnđ
* Số lượng sản phẩm giá tiền dưới 1.000.000 vnđ chiếm số lượng lớn (7467 sản phẩm)
* Số lượng sản phẩm theo có xu hướng tỉ lệ nghịch với khoảng giá sản phẩm

=> Sử dụng khoảng giá dưới 5.000.000 để lập biểu đồ thể hiện xu hướng số lượng sản phẩm theo giá, tránh việc đồ thị quá dài, nhỏ.



* So sánh tương quan giữa “original\_price” và “price”

=> Có sự khách biệt về khoảng cách của 2 loại giá cả trong các sản phẩm khác nhau. Nhưng xu hướng chung là khoảng cách không quá lớn. Nếu khoảng cách quá lớn có thể sử dụng dữ liệu đó để xem xét ncc có đang bán phá giá, khai khống giá để gian lận về việc đẩy sản phẩm lên gian “SALE”

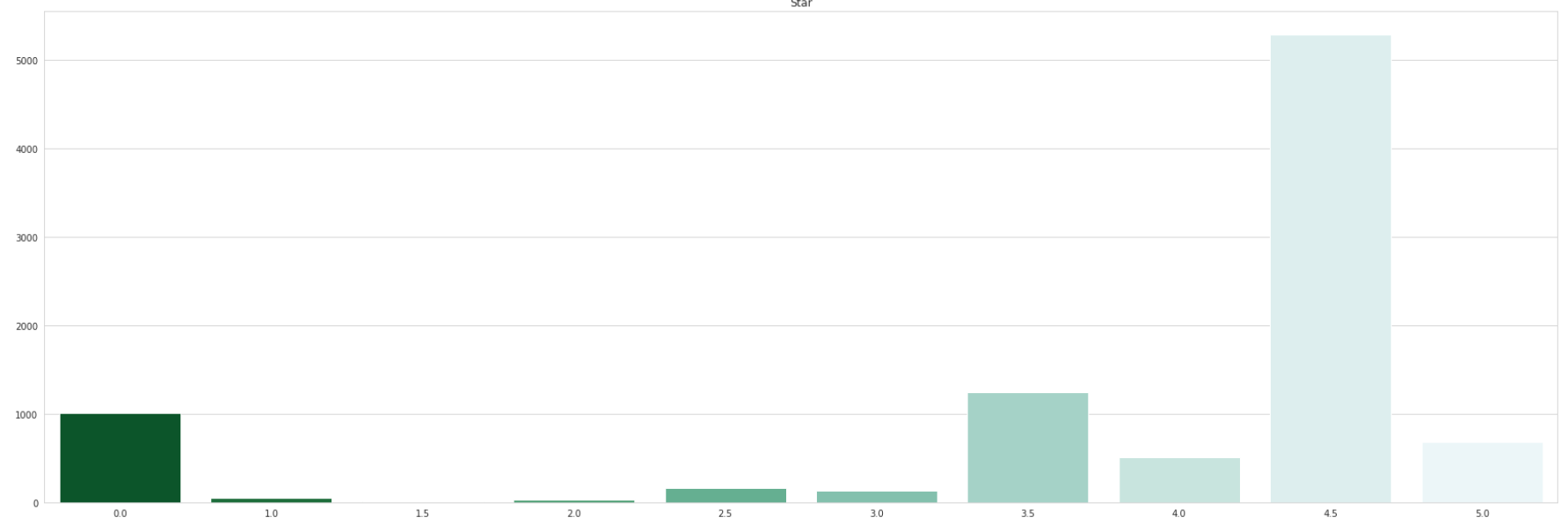
=> Xuất hiện các điểm kì dị (VD: “price” lớn hơn “original\_price”). Cần làm sạch dữ liệu lần 2 (loại bỏ các khoảng giá trị bị miss) nếu dùng để thực hiện khảo sát trên

### Đánh giá nhà cung cấp các sản phẩm



* Nhận xét được Tiki Trading có số lượng sản phẩm nhiều nhất do Tiki phát triển từ nhà cung cấp chính Tiki.
* Lập top các nhà cung cấp có nhiều sản phẩm nhất tính từ sau Tiki Trading để đánh giá sự đóng góp của các ncc với sàn TMĐT này.

### Đánh giá chất lượng sản phẩm



* Nhìn nhận thấy số lượng sản phẩm được đánh giá chủ yếu từ mức 3 sao đổ lên (Nhiều nhất là 4 sao)

=> Chất lượng sản phẩm trên Tiki mặt bằng chung tốt, là sàn TMĐT kiểm soát chất lượng sản phẩm

* Đặt cho sản phẩm 1 số điểm theo công thức score = số đánh giá \* số sao trung bình (do Tiki không hiển thị số lượng sản phẩm bán ra nên ta coi số đánh giá tương ứng với số lượng bán được)

=> Xếp hạng được sản phẩm, thể hiện xu hướng mua hàng, chất lượng chung. Đánh giá tương quan giữa số điểm với số sao/số lượng đánh giá:

* VD: Cùng 1 số điểm tương đương nhau nhưng 1 bên số sao trung bình cao hơn, rút được chất lượng sản phẩm tốt nhưng chưa tiếp cận được với nhiều khách hàng => đẩy mạnh tiếp thị

### Đánh giá theo độ chi tiết của thông tin sản phẩm

* Chuyển phần text của các mục “brief”, “description”, “info” thành số ký tự để đánh giá độ chi tiết của phần mô tả
* So sánh ví dụ tương quan giữa “price” và “info”

=> Đánh giá chưa thực sự hiệu quả nhưng có thể nhận thấy các sản phẩm có giá tầm trung thường có phần giới thiệu dài, chi tiết