Bài 1
Tổng quan về phân tích dữ liệu

Khóa học: Phân tích dữ liệu với Python

Mục tiêu

- Trình bày được quy trình phân tích dữ liệu
- Trình bày được mối quan hệ giữa phân tích dữ liệu và khoa học dữ liệu
- Phân biệt được dữ liệu định tính và dữ liệu định lượng
- Xác định được thang đo dữ liệu cho từng thuộc tính

Thảo luận

- Trước khi ra quyết định chúng ta cần làm gì

Bác nông dân đang nghĩ gì?

- Mình đang nuôi vịt thành công?
- Mình có bao nhiêu con vịt?
- Hôm nay số vịt của mình vẫn giữ nguyên?
- Các con vịt của mình đang lớn lên, hay còi đi?
- …

- Nếu như có một trang trại hàng triệu con vịt, người nông dân còn cần biết những điều trên?

Bạn muốn tìm một công việc?

- Search tin tuyển dụng
- So sánh khả năng bản thân ~ yêu cầu công việc
- So sánh mức lương giữ các tin tuyển dụng
- ….?

Chủ của một xưởng sản xuất muốn biết gì?

- Mọi thứ đề ổn?
  - Khả năng cung cấp > nhu cầu đặt hàng?
  - Doanh thu > chi phí?
  - Doanh thu năm nay cao hơn năm trước?
  - Chiến lược sản xuất phù hợp với nhu cầu thị trường?
  - … ?

Nội dung

1. Phân tích dữ liệu
2. Quy trình phân tích dữ liệu
3. Các thang đo dữ liệu
4. Demo

1. Phân tích dữ liệu

Phân tích dữ liệu là một quy trình thu thập, làm sạch, biến đổi, mô hình hóa dữ liệu với mục tiêu tìm kiếm những thông tin hữu ích, đề xuất những kết luận và hỗ trợ ra quyết định

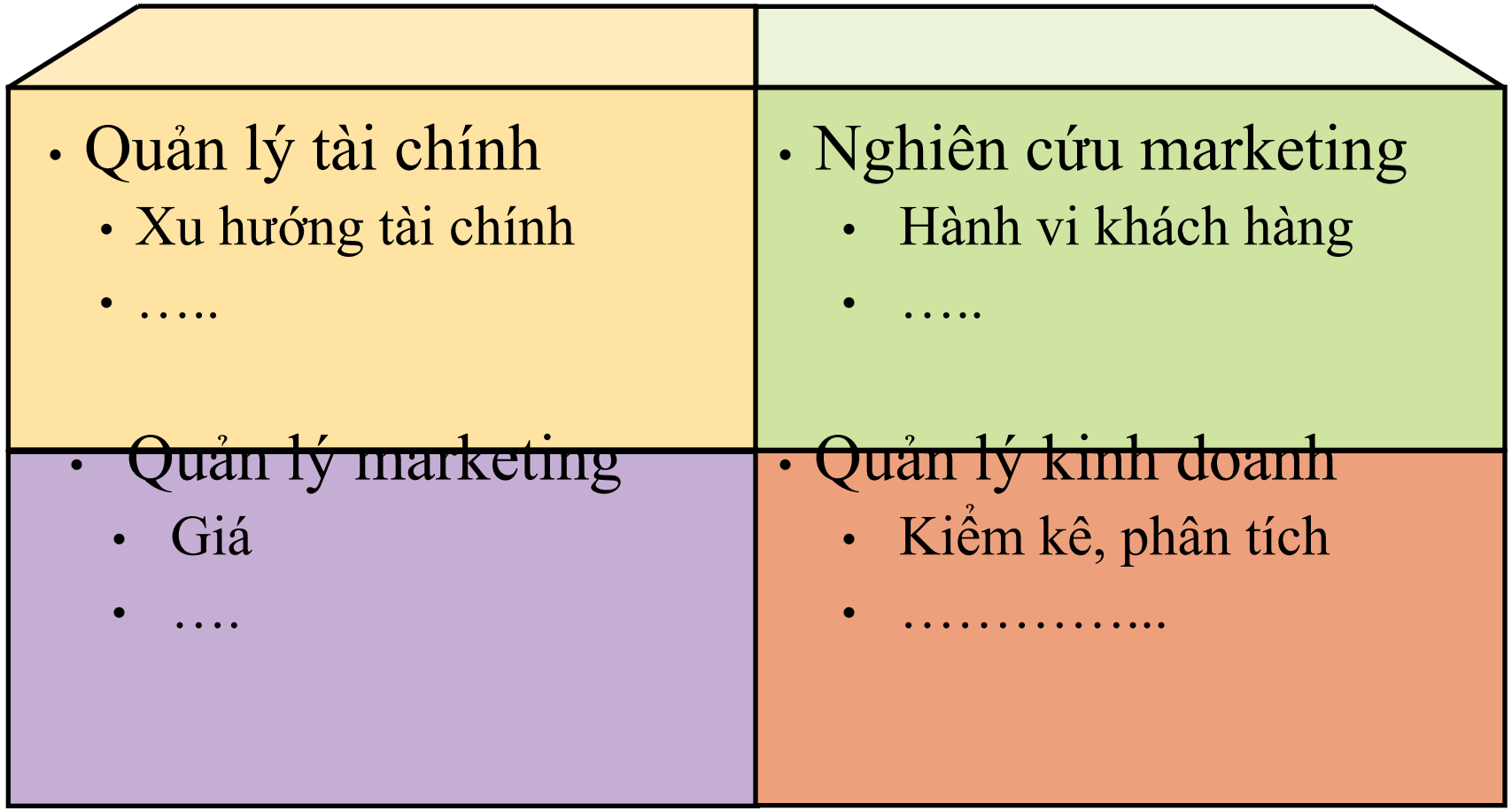
Thời kỳ Chiếm hữu nô lệ
- Ghi chép

Thời kỳ Phong kiến
- Phân tích theo không gian & thời gian

Thời kỳ SX hàng hóa
- Thể hiện quan hệ giữa lượng và chất

Ngày nay
- Là một trong những công cụ quản lý

Phân tích dữ liệu với các ngành khác



Phân tích dữ liệu – Khoa học dữ liệu

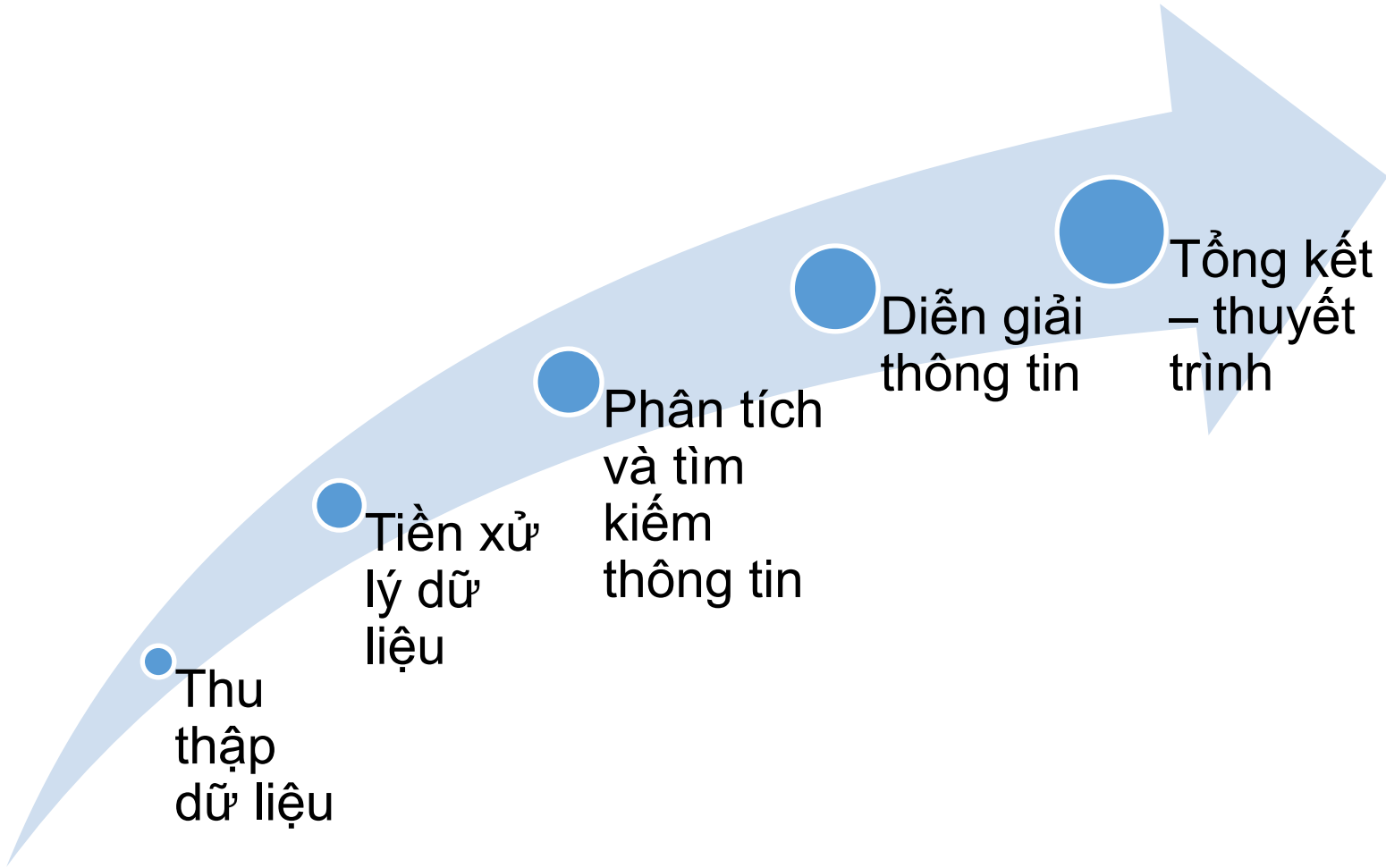
Phân tích dữ liệu

- Phát hiện thông tin từ dữ liệu
- Trả lời các câu hỏi của nhà quản lý

Khoa học dữ liệu

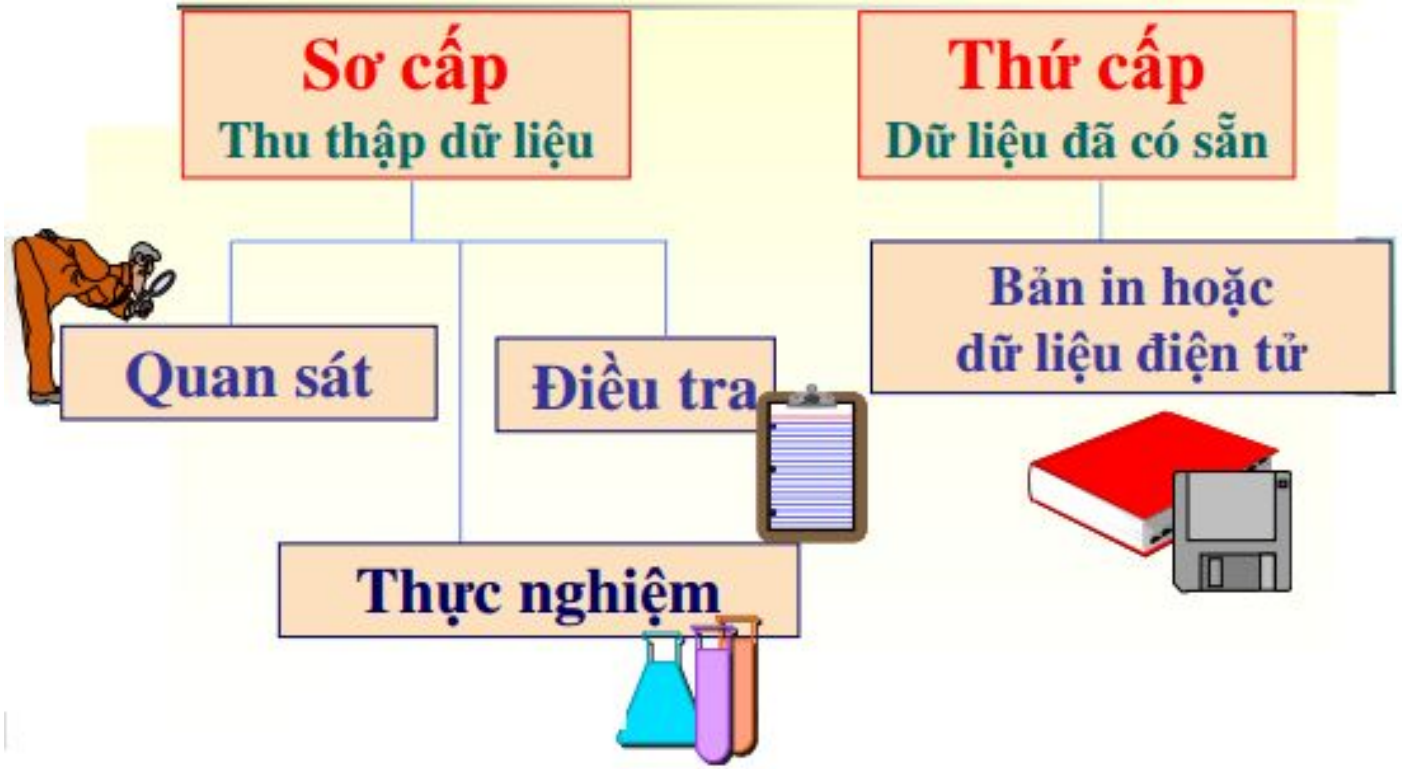
- Dự đoán tương lai từ dữ liệu
- Thường tự đề xuất vấn đề cần giải quyết

2. Quy trình phân tích dữ liệu



2.1 Thu thập dữ liệu

- Yêu cầu: luôn phải hiểu bài toán trước khi tiến hành thu thập dữ liệu

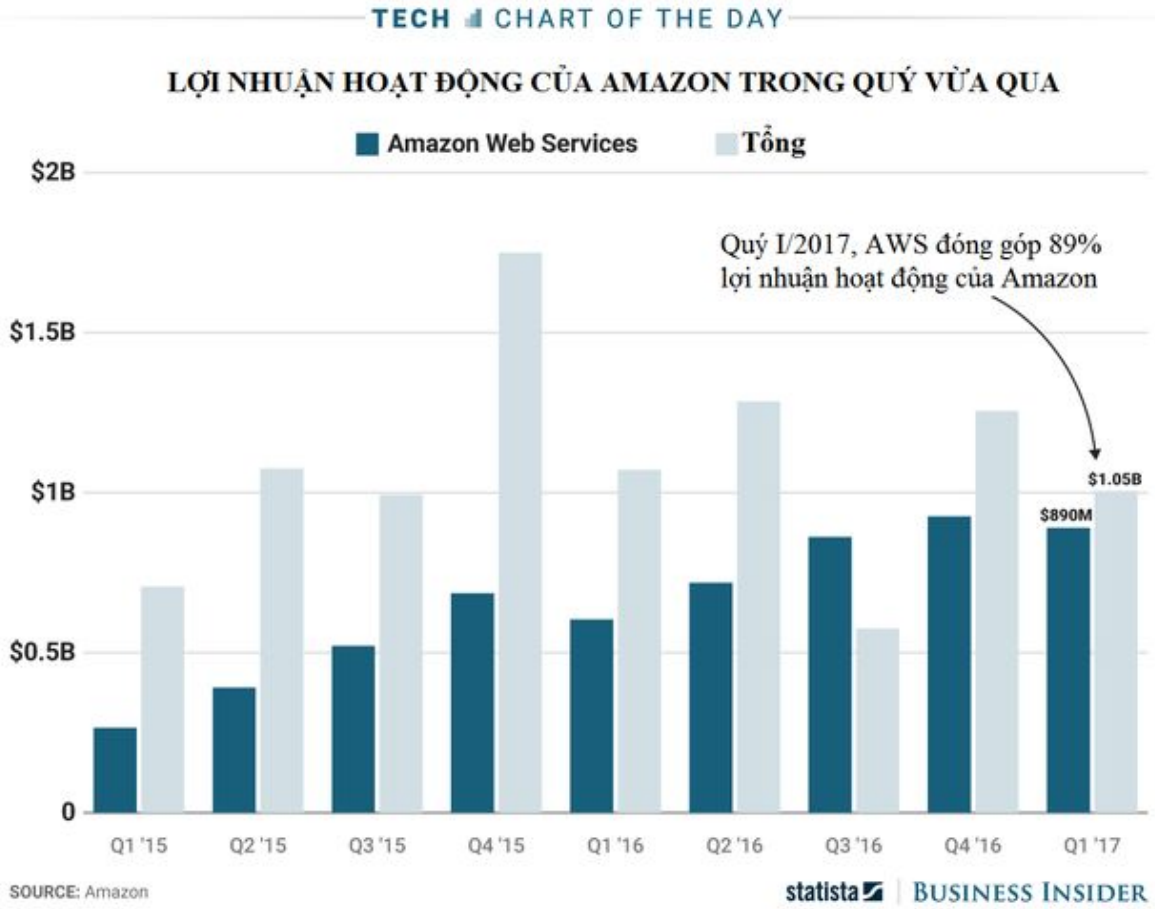


Tiền xử lý dữ liệu

- Lọc dữ liệu
- Làm sạch dữ liệu
- Biến đổi dữ liệu

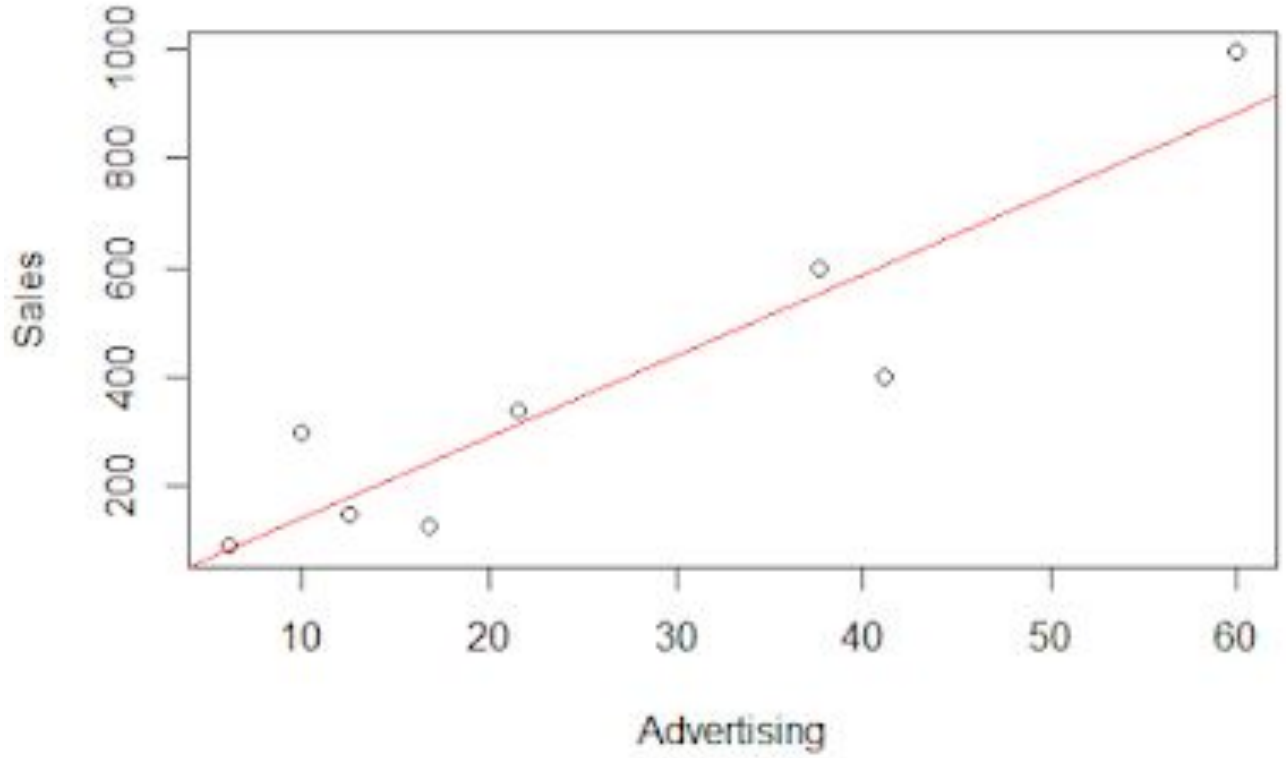
Phân tích và tìm kiếm thông tin

- Mô tả dữ liệu
- Phân tích phân bố dữ liệu
- Trực quan hóa dữ liệu
- Kiểm định giả thiết



Diễn giải thông tin

- Hiểu thông tin chi tiết
- Tìm được tác động của các yếu tố lên hệ thống



3. Các thang đo dữ liệu

4. Demo

- Môi trường cần chuẩn bị:
  - Python 3
  - Jupyter notebook
  - Cài đặt thư viện: pandas, matplotlib (đã có sẵn nếu sử dụng anaconda)

Phân tích bộ dữ liệu Online Retail

- Phân tích tình hình kinh doanh dựa trên thông tin bán hàng được cung cấp trong bảng dữ liệu
  - InvoiceNo: Số hóa đơn
  - StockCode: mã hàng
  - Description: Mô tả hàng
  - Quantity: Số lương
  - InvoiceDate: Ngày bán
  - UnitPrice: Đơn giá
  - CustomerID: Mã khách
  - Country: Nước sản xuất

Chủ doanh nghiệp muốn biết

- Công ty bán hàng do bao nhiêu nước sản xuất
- Tổng số lượng đơn hàng bán ra, tổng doanh thu
- Top 10 mặt hàng có số lượng bán ra lớn nhất
- Top 10 mặt hàng có doanh thu lớn nhất

Phân tích

- Cần phải hiểu rõ bộ dữ liệu trước khi trả lời câu hỏi

```
import pandas as pd
data=pd.read_csv('data\OnlineRetail.csv', encoding = "ISO-8859-1")
# hiển thị 5 dòng dữ liệu đầu tiên
data.head()
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

Phân tích (tiếp)

- Tìm hiểu cấu trúc bộ dữ liệu

```
print (data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  object
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: float64(2), int64(1), object(5)
memory usage: 33.1+ MB
None
```

Phân tích (tiếp)

- Công ty bán hàng của bao nhiêu quốc gia

```
# lấy ra tên các quốc gia
countries = data.Country.unique()
print ("số lượng các quốc gia: " + str(countries.size))
```

Phân tích (tiếp)

- ***Số lượng đơn hàng bán ra và tổng doanh thu***

```
# Tạo cột tính thành tiền của các mặt hàng
data['total'] = data['Quantity'] * data['UnitPrice']

# Giá trị đơn hàng của mỗi đơn hàng
total_invoices = data['total'].sum()
print ("số lượng hóa đơn bán ra: "+ str (total_invoices.size))
print ("Tổng doanh thu: " + str(total_invoices.sum()))
```

Phân tích (tiếp)

- ***Top 10 mặt hàng có số lượng bán ra lớn rất***

```
quantity_product = data.groupby(['StockCode’,
'Description'])['Quantity'].sum().sort_values(ascending= False)
quantity_product.head(10)
```

Phân tích (tiếp)

- ***Top 10 mặt hàng có doanh thu lớn nhất***

```
quantity_product = data.groupby(['StockCode’,
'Description'])['total'].sum().sort_values(ascending= False)
quantity_product.head(10)
```

Tóm tắt bài học

- Phân tích dữ liệu là một công cụ hỗ trợ việc ra quyết định
- Quy trình phân tích dữ liệu gồm 5 bước: Thu thập dữ liệu, tiền xử lý dữ liệu, phân tích và tìm kiếm thông tin, diễn giải thông tin, tổng kết - thuyết trình
- Phân tích dữ liệu phát hiện thông tin từ dữ liệu
- Khoa học dữ liệu dự đoán tương lai từ dữ liệu
- Thuộc tính định tính: giá trị phụ thuộc vào đánh giá của con người
- Thuộc tính định lượng: do cân, đo, đong, đếm mà có được
- Các loại thang đo: Nominal, Ordinal, Interval, Ratio