

ĐỌC DỮ LIỆU

In [2]: *# Đọc dữ liệu*

```
import pandas as pd
import numpy as np
```

```
df = pd.read_csv('../data/orginal_sales_data_edit.csv', encoding='utf-8', header=0, delimiter=',')
```

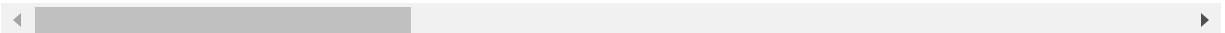
HIỂN THỊ BẢNG

In [3]: *# Hiển thị thông tin bảng*
df

Out[3]:

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERI
0	10107	30	95.70	2	2871.00	2/24
1	10121	34	81.35	5	2765.90	5/7
2	10134	41	94.74	2	3884.34	7/1
3	10145	45	83.26	6	3746.70	8/25
4	10159	49	100.00	14	5205.27	10/10
...
2818	10350	20	100.00	15	2244.40	12/2
2819	10373	29	100.00	1	3978.51	1/31
2820	10386	43	100.00	4	5417.57	3/1
2821	10397	34	62.24	1	2116.16	3/28
2822	10414	47	65.52	9	3079.44	5/6

2823 rows × 28 columns

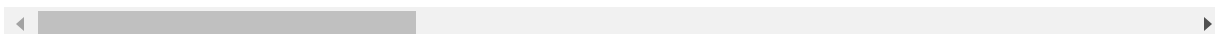


In [4]: *# Đọc 10 dòng đầu*
df.head(10)

Out[4]:

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDAT
0	10107	30	95.70	2	2871.00	2/24/200
1	10121	34	81.35	5	2765.90	5/7/200
2	10134	41	94.74	2	3884.34	7/1/200
3	10145	45	83.26	6	3746.70	8/25/200
4	10159	49	100.00	14	5205.27	10/10/200
5	10168	36	96.66	1	3479.76	10/28/200
6	10180	29	86.13	9	2497.77	11/11/200
7	10188	48	100.00	1	5512.32	11/18/200
8	10201	22	98.57	2	2168.54	12/1/200
9	10211	41	100.00	14	4708.44	1/15/200

10 rows × 28 columns

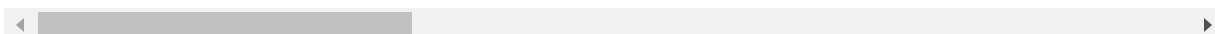


In [5]: *# Đọc 10 dòng cuối*
df.tail(10)

Out[5]:

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERI
2813	10293	32	60.06	1	1921.92	9/9
2814	10306	35	59.51	6	2082.85	10/14
2815	10315	40	55.69	5	2227.60	10/29
2816	10327	37	86.74	4	3209.38	11/10
2817	10337	42	97.16	5	4080.72	11/21
2818	10350	20	100.00	15	2244.40	12/2
2819	10373	29	100.00	1	3978.51	1/31
2820	10386	43	100.00	4	5417.57	3/1
2821	10397	34	62.24	1	2116.16	3/28
2822	10414	47	65.52	9	3079.44	5/6

10 rows × 28 columns



```
In [6]: # tạo ra 1 tóm tắt thống kê cho dữ liệu DataFrame
        """
        df.describe() sẽ tính toán các thống kê quan trọng cho từng cột trong DataFrame
        """

        df.describe()
```

Out[6]:

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	
count	2823.000000	2823.000000	2823.000000	2823.000000	2823.000000	2
mean	10258.725115	35.092809	83.658544	6.466171	3553.889072	
std	92.085478	9.741443	20.174277	4.225841	1841.865106	
min	10100.000000	6.000000	26.880000	1.000000	482.130000	
25%	10180.000000	27.000000	68.860000	3.000000	2203.430000	
50%	10262.000000	35.000000	95.700000	6.000000	3184.800000	
75%	10333.500000	43.000000	100.000000	9.000000	4508.000000	
max	10425.000000	97.000000	100.000000	18.000000	14082.800000	

TỔNG QUAN DỮ LIỆU

```
In [7]: """
Xem thông tin định dạng và số lượng quan sát non-null của mỗi trường trong DataFrame
"""

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2823 entries, 0 to 2822
Data columns (total 28 columns):
#   Column                Non-Null Count  Dtype
---  -
0   ORDERNUMBER           2823 non-null  int64
1   QUANTITYORDERED       2823 non-null  int64
2   PRICEEACH             2823 non-null  float64
3   ORDERLINENUMBER       2823 non-null  int64
4   SALES                 2823 non-null  float64
5   ORDERDATE             2823 non-null  object
6   STATUS                2823 non-null  object
7   QTR_ID                2823 non-null  int64
8   MONTH_ID              2823 non-null  int64
9   YEAR_ID               2823 non-null  int64
10  PRODUCTLINE           2823 non-null  object
11  MSRP                  2823 non-null  int64
12  PRODUCTCODE           2823 non-null  object
13  CATEGORY              2823 non-null  object
14  SUBCATEGORY           2823 non-null  object
15  CUSTOMERNAME          2823 non-null  object
16  PHONE                 2823 non-null  object
17  ADDRESSLINE1           2823 non-null  object
18  ADDRESSLINE2           302 non-null   object
19  CITY                  2823 non-null  object
20  STATE                 1337 non-null  object
21  POSTALCODE            2747 non-null  object
22  COUNTRY               2823 non-null  object
23  TERRITORY             1749 non-null  object
24  CONTACTLASTNAME       2823 non-null  object
25  CONTACTFIRSTNAME      2823 non-null  object
26  DEALSIZE              2823 non-null  object
27  PAYMENTFULLNAME       2823 non-null  object
dtypes: float64(2), int64(7), object(19)
memory usage: 617.7+ KB
```

KIỂM TRA CẤU TRÚC CHỈ MỤC DỮ LIỆU

```
In [8]: # Kiểm tra cấu trúc chỉ mục dữ liệu

df.index
```

```
Out[8]: RangeIndex(start=0, stop=2823, step=1)
```

KIỂM TRA BAO NHIÊU DÒNG, CỘT

In [9]: *# Kt bnhiu dòng, cột*

```
df.shape
```

```
""""  
28 cột  
2823 dòng  
""""
```

Out[9]: `\n28 cột\n2823 dòng\n`

KIỂM TRA KIỂU DỮ LIỆU

In [10]: *# Kt kiểu dữ liệu của cột YEAR_ID*

```
df["YEAR_ID"].dtype
```

Out[10]: `dtype('int64')`

CHUYỂN ĐỔI DỮ LIỆU TỪ int64 SANG int32

In [11]: *# Chuyển đổi dữ liệu từ int64 sang int32*

```
df["YEAR_ID"] = df["YEAR_ID"].astype("int32")
```

In [12]: *# Kt kiểu dữ liệu của cột YEAR_ID*

```
df["YEAR_ID"].dtype
```

Out[12]: `dtype('int32')`

LOẠI BỎ CÁC DÒNG CÓ GIÁ TRỊ NaN

```
In [13]: # Loại bỏ các dòng có giá trị NaN

df.dropna(how='all', inplace=True)

# Hiển thị từ dòng đầu tiên đến dòng thứ 10
df[:10]
```

```
Out[13]:
```

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE
0	10107	30	95.70	2	2871.00	2/24/2001
1	10121	34	81.35	5	2765.90	5/7/2001
2	10134	41	94.74	2	3884.34	7/1/2001
3	10145	45	83.26	6	3746.70	8/25/2001
4	10159	49	100.00	14	5205.27	10/10/2001
5	10168	36	96.66	1	3479.76	10/28/2001
6	10180	29	86.13	9	2497.77	11/11/2001
7	10188	48	100.00	1	5512.32	11/18/2001
8	10201	22	98.57	2	2168.54	12/1/2001
9	10211	41	100.00	14	4708.44	1/15/2002

10 rows × 6 columns

LOẠI BỎ CÁC DÒNG CÓ GIÁ TRỊ TRÙNG LẶP

```
In [14]: # Loại bỏ các dòng có giá trị trùng lặp

df.drop_duplicates(inplace=True)
```

TÍNH SỐ LƯỢNG CÁC DÒNG CÓ GIÁ TRỊ NaN TRONG CỘT "ADDRESSLINE2"

```
In [15]: # Tính số lượng các dòng có giá trị NaN trong cột "ADDRESSLINE2"

df["ADDRESSLINE2"].isna().sum()
```

Out[15]: 2521

THAY THẾ TẤT CẢ CÁC DÒNG CÓ GIÁ TRỊ NaN TRONG CỘT "ADDRESSLINE2" BẰNG CHUỖI 'Unknown'

In [20]: *# Thay thế tất cả các dòng có giá trị NaN trong cột "ADDRESSLINE2" bằng chuỗi 'Unknown'*

```
df["ADDRESSLINE2"].fillna('Unknown', inplace=True)  
df["ADDRESSLINE2"]
```

DẠNG SERIES

In [21]: *# Lấy dữ liệu theo cột dạng chuỗi*

```
df["QUANTITYORDERED"]
```

Out[21]:

0	30
1	34
2	41
3	45
4	49
..	
2818	20
2819	29
2820	43
2821	34
2822	47

Name: QUANTITYORDERED, Length: 2823, dtype: int64

DẠNG MẢNG

In [18]: *# Lấy dữ liệu về 1 mảng*

```
df["QUANTITYORDERED"].values
```

Out[18]: array([30, 34, 41, ..., 43, 34, 47], dtype=int64)

DẠNG DATAFRAME

```
In [23]: df[['QUANTITYORDERED']]
```

```
Out[23]:
```

	QUANTITYORDERED
0	30
1	34
2	41
3	45
4	49
...	...
2818	20
2819	29
2820	43
2821	34
2822	47

2823 rows × 1 columns

GỘP CỘT TRẢ VỀ DẠNG DATAFRAME

```
In [24]: df[['QUANTITYORDERED', 'ORDERNUMBER', 'PRICEEACH']]
```

```
Out[24]:
```

	QUANTITYORDERED	ORDERNUMBER	PRICEEACH
0	30	10107	95.70
1	34	10121	81.35
2	41	10134	94.74
3	45	10145	83.26
4	49	10159	100.00
...
2818	20	10350	100.00
2819	29	10373	100.00
2820	43	10386	100.00
2821	34	10397	62.24
2822	47	10414	65.52

2823 rows × 3 columns

In [25]: `df[['QUANTITYORDERED', 'PRICEEACH', 'SALES', 'MSRP']]`

Out[25]:

	QUANTITYORDERED	PRICEEACH	SALES	MSRP
0	30	95.70	2871.00	95
1	34	81.35	2765.90	95
2	41	94.74	3884.34	95
3	45	83.26	3746.70	95
4	49	100.00	5205.27	95
...
2818	20	100.00	2244.40	54
2819	29	100.00	3978.51	54
2820	43	100.00	5417.57	54
2821	34	62.24	2116.16	54
2822	47	65.52	3079.44	54

2823 rows × 4 columns

TẠO CỘT MỚI VÀ TRẢ VỀ BẢNG DATAFRAME

In [27]: `# Tạo cột mới và trả về bảng DataFrame`
`# Tạo 1 cột mới có tên là TOTALPRICE`
`# Giá trị TOTALPRICE = Giá trị QUANTITYORDERED * Giá trị PRICEEACH`
`df['TOTALPRICE'] = df['QUANTITYORDERED'] * df['PRICEEACH']`

In [29]: *# Kiểm tra cột TOTALPRICE sau khi tạo mới*

```
df[['TOTALPRICE']]
```

Out[29]:

	TOTALPRICE
0	2871.00
1	2765.90
2	3884.34
3	3746.70
4	4900.00
...	...
2818	2000.00
2819	2900.00
2820	4300.00
2821	2116.16
2822	3079.44

2823 rows × 1 columns

In []: *# Tạo 1 cột mới có tên là FULLNAME*

Dữ liệu "FULLNAME" = "CONTACTLASTNAME" + " " + "CONTACTFIRSTNAME"

In [30]: `df['FULLNAME'] = df['CONTACTLASTNAME'] + ' ' + df['CONTACTFIRSTNAME']`

In [32]: *# Kiểm tra cột FULLNAME sau khi tạo*

```
df[['FULLNAME']]
```

Out[32]:

	FULLNAME
0	Yu Kwai
1	Henriot Paul
2	Da Cunha Daniel
3	Young Julie
4	Brown Julie
...	...
2818	Freyre Diego
2819	Koskitalo Pirkko
2820	Freyre Diego
2821	Roulet Annette
2822	Yoshido Juri

2823 rows × 1 columns

LẤY DỮ LIỆU THEO DÒNG

```
In [33]: df[4:10]
```

Out[33]:

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE
4	10159	49	100.00	14	5205.27	10/10/2003
5	10168	36	96.66	1	3479.76	10/28/2003
6	10180	29	86.13	9	2497.77	11/11/2003
7	10188	48	100.00	1	5512.32	11/18/2003
8	10201	22	98.57	2	2168.54	12/1/2003
9	10211	41	100.00	14	4708.44	1/15/2004

6 rows × 7 columns

