การจัดการข้อมูลสูญหาย

การรวบรวมข้อมูลมาวิเคราะห์นั้น บางครั้งอาจจะมีข้อมูลที่ได้มา ไม่ครบบ้าง ตกหล่นหรือขาดหายไปบ้างเรียกส่วนนี้ว่า Missing Data หรือ Missing Value ในหัวข้อนี้จะมาตรวจสอบข้อมูลและจัดการข้อมูลสูญหาย (Clean Data)

In [1]:
```
import pandas as pd

df = pd.read_csv("datasets/Employee.csv")
df
```

Out[1]:

	Name	Job	Age	Salary	Bonus	Address
0	A	Programmer	20.0	30000.0	10%	123.0
1	B	Programmer	18.0	NaN	10%	NaN
2	C	Developer	NaN	32000.0	10%	NaN
3	D	NaN	23.0	40000.0	10%	NaN
4	E	Sale	29.0	40000.0	10%	NaN
5	F	Manager	NaN	75000.0	10%	NaN
6	NaN	NaN	NaN	NaN	NaN	NaN
7	H	Maketing	34.0	60000.0	10%	NaN
8	I	NaN	26.0	100000.0	10%	NaN
9	H	Maketing	34.0	60000.0	10%	NaN
10	E	Sale	29.0	40000.0	10%	NaN

In [2]:
```
import pandas as pd

df = pd.read_csv("datasets/Employee.csv", index_col="Name")
df
```

Out[2]:

Name	Job	Age	Salary	Bonus	Address
A	Programmer	20.0	30000.0	10%	123.0
B	Programmer	18.0	NaN	10%	NaN
C	Developer	NaN	32000.0	10%	NaN
D	NaN	23.0	40000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN
F	Manager	NaN	75000.0	10%	NaN
NaN	NaN	NaN	NaN	NaN	NaN
H	Maketing	34.0	60000.0	10%	NaN
I	NaN	26.0	100000.0	10%	NaN
H	Maketing	34.0	60000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN

In [3]:
```
df.shape
```

Out[3]:
```
(11, 5)
```

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 11 entries, A to E
Data columns (total 5 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   Job      8 non-null      object 
 1   Age      8 non-null      float64
 2   Salary   9 non-null      float64
 3   Bonus    10 non-null     object 
 4   Address  1 non-null      float64
dtypes: float64(3), object(2)
memory usage: 528.0+ bytes
```

In [5]:
```
#การตรวจสอบข้อมูลสูญหายด้วย isnull()
df.isnull()
```

Out[5]:

Name	Job	Age	Salary	Bonus	Address
A	False	False	False	False	False
B	False	False	True	False	True
C	False	True	False	False	True
D	True	False	False	False	True
E	False	False	False	False	True
F	False	True	False	False	True
NaN	True	True	True	True	True
H	False	False	False	False	True
I	True	False	False	False	True
H	False	False	False	False	True
E	False	False	False	False	True

In [6]:
```
#ตรวจสอบว่ามีคอลัมน์ใดบ้างที่ไม่มีข้อมูล
df.isnull().any()
```

Out[6]:
```
Job        True
Age        True
Salary     True
Bonus      True
Address    True
dtype: bool
```

In [7]:
```
#นับจำนวนคอลัมน์ที่ไม่มีข้อมูล
df.isnull().sum()
```

Out[7]:
```
Job         3
Age         3
Salary      2
Bonus       1
Address    10
dtype: int64
```

In [8]:
```
#การตรวจสอบข้อมูลครบถ้วนด้วย notnull()
df.notnull()
```

Out[8]:

Name	Job	Age	Salary	Bonus	Address
A	True	True	True	True	True
B	True	True	False	True	False
C	True	False	True	True	False
D	False	True	True	True	False
E	True	True	True	True	False
F	True	False	True	True	False
NaN	False	False	False	False	False
H	True	True	True	True	False
I	False	True	True	True	False
H	True	True	True	True	False
E	True	True	True	True	False

In [9]:
```
#ตรวจสอบว่ามีคอลัมน์ใดบ้างที่มีข้อมูล
df.notnull().any()
```

Out[9]:
```
Job        True
Age        True
Salary     True
Bonus      True
Address    True
dtype: bool
```

In [10]:
```
#นับจำนวนคอลัมน์ที่มีข้อมูล
df.isnull().sum()
```

Out[10]:
```
Job         3
Age         3
Salary      2
Bonus       1
Address    10
dtype: int64
```

วิธีจัดการข้อมูลสูญหาย

- แทนที่ด้วยค่าเฉลี่ยข้อมูลทั้งหมด
- แทนที่ด้วยค่าตรงๆที่กำหนดขึ้นมา
- แทนที่ด้วยค่าก่อนหน้า
- แทนที่ด้วยค่าถัดไป
- ลบข้อมูล

แทนที่ด้วยค่าเฉลี่ยข้อมูลทั้งหมด

In [11]:
```
df.describe()
```

Out[11]:

	Age	Salary	Address
count	8.000000	9.000000	1.0
mean	26.625000	53000.000000	123.0
std	5.998512	23097.618925	NaN
min	18.000000	30000.000000	123.0
25%	22.250000	40000.000000	123.0
50%	27.500000	40000.000000	123.0
75%	30.250000	60000.000000	123.0
max	34.000000	100000.000000	123.0

In [12]:
```
#นำเข้า DataFrame ใหม่
df = pd.read_csv("datasets/Employee.csv", index_col="Name")
df
```

Out[12]:

Name	Job	Age	Salary	Bonus	Address
A	Programmer	20.0	30000.0	10%	123.0
B	Programmer	18.0	NaN	10%	NaN
C	Developer	NaN	32000.0	10%	NaN
D	NaN	23.0	40000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN
F	Manager	NaN	75000.0	10%	NaN
NaN	NaN	NaN	NaN	NaN	NaN
H	Maketing	34.0	60000.0	10%	NaN
I	NaN	26.0	100000.0	10%	NaN
H	Maketing	34.0	60000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN

In [13]:
```
#แทนที่ด้วยค่าเฉลี่ยข้อมูลทั้งหมด
df['Salary'] = df['Salary'].fillna(df['Salary'].mean())
df
```

Out[13]:

Name	Job	Age	Salary	Bonus	Address
A	Programmer	20.0	30000.0	10%	123.0
B	Programmer	18.0	53000.0	10%	NaN
C	Developer	NaN	32000.0	10%	NaN
D	NaN	23.0	40000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN
F	Manager	NaN	75000.0	10%	NaN
NaN	NaN	NaN	53000.0	NaN	NaN
H	Maketing	34.0	60000.0	10%	NaN
I	NaN	26.0	100000.0	10%	NaN
H	Maketing	34.0	60000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN

แทนที่ด้วยค่าตรงๆที่กำหนดขึ้นมา

In [14]:
```
#นำเข้า DataFrame ใหม่
df = pd.read_csv("datasets/Employee.csv", index_col="Name")
df
```

Out[14]:

Name	Job	Age	Salary	Bonus	Address
A	Programmer	20.0	30000.0	10%	123.0
B	Programmer	18.0	NaN	10%	NaN
C	Developer	NaN	32000.0	10%	NaN
D	NaN	23.0	40000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN
F	Manager	NaN	75000.0	10%	NaN
NaN	NaN	NaN	NaN	NaN	NaN
H	Maketing	34.0	60000.0	10%	NaN
I	NaN	26.0	100000.0	10%	NaN
H	Maketing	34.0	60000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN

In [15]:
```
#แทนที่ด้วยค่าตรงๆที่กำหนดขึ้นมา
df['Salary'] = df['Salary'].fillna(22000)
df
```

Out[15]:

Name	Job	Age	Salary	Bonus	Address
A	Programmer	20.0	30000.0	10%	123.0
B	Programmer	18.0	22000.0	10%	NaN
C	Developer	NaN	32000.0	10%	NaN
D	NaN	23.0	40000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN
F	Manager	NaN	75000.0	10%	NaN
NaN	NaN	NaN	22000.0	NaN	NaN
H	Maketing	34.0	60000.0	10%	NaN
I	NaN	26.0	100000.0	10%	NaN
H	Maketing	34.0	60000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN

แทนที่ด้วยค่าก่อนหน้า

In [16]:
```
#นำเข้า DataFrame ใหม่
df = pd.read_csv("datasets/Employee.csv", index_col="Name")
df
```

Out[16]:

Name	Job	Age	Salary	Bonus	Address
A	Programmer	20.0	30000.0	10%	123.0
B	Programmer	18.0	NaN	10%	NaN
C	Developer	NaN	32000.0	10%	NaN
D	NaN	23.0	40000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN
F	Manager	NaN	75000.0	10%	NaN
NaN	NaN	NaN	NaN	NaN	NaN
H	Maketing	34.0	60000.0	10%	NaN
I	NaN	26.0	100000.0	10%	NaN
H	Maketing	34.0	60000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN

In [17]:
```
#แทนที่ด้วยค่าก่อนหน้า
df.fillna(method='pad')
```

```
/var/folders/83/3fg00w111r7bf7rcsz4nznlh0000gn/T/ipykernel_21816/3352054385.
py:2: FutureWarning: DataFrame.fillna with 'method' is deprecated and will r
aise in a future version. Use obj.ffill() or obj.bfill() instead.
  df.fillna(method='pad')
```

Out[17]:

Name	Job	Age	Salary	Bonus	Address
A	Programmer	20.0	30000.0	10%	123.0
B	Programmer	18.0	30000.0	10%	123.0
C	Developer	18.0	32000.0	10%	123.0
D	Developer	23.0	40000.0	10%	123.0
E	Sale	29.0	40000.0	10%	123.0
F	Manager	29.0	75000.0	10%	123.0
NaN	Manager	29.0	75000.0	10%	123.0
H	Maketing	34.0	60000.0	10%	123.0
I	Maketing	26.0	100000.0	10%	123.0
H	Maketing	34.0	60000.0	10%	123.0
E	Sale	29.0	40000.0	10%	123.0

แทนที่ด้วยค่าถัดไป

In [18]:
```
#นำเข้า DataFrame ใหม่
df = pd.read_csv("datasets/Employee.csv", index_col="Name")
df
```

Out[18]:

Name	Job	Age	Salary	Bonus	Address
A	Programmer	20.0	30000.0	10%	123.0
B	Programmer	18.0	NaN	10%	NaN
C	Developer	NaN	32000.0	10%	NaN
D	NaN	23.0	40000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN
F	Manager	NaN	75000.0	10%	NaN
NaN	NaN	NaN	NaN	NaN	NaN
H	Maketing	34.0	60000.0	10%	NaN
I	NaN	26.0	100000.0	10%	NaN
H	Maketing	34.0	60000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN

In [19]:
```
#แทนที่ด้วยค่าถัดไป
df.fillna(method='bfill')
```

```
/var/folders/83/3fg00w111r7bf7rcsz4nznlh0000gn/T/ipykernel_21816/2964409054.
py:2: FutureWarning: DataFrame.fillna with 'method' is deprecated and will r
aise in a future version. Use obj.ffill() or obj.bfill() instead.
  df.fillna(method='bfill')
```

Out[19]:

Name	Job	Age	Salary	Bonus	Address
A	Programmer	20.0	30000.0	10%	123.0
B	Programmer	18.0	32000.0	10%	NaN
C	Developer	23.0	32000.0	10%	NaN
D	Sale	23.0	40000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN
F	Manager	34.0	75000.0	10%	NaN
NaN	Maketing	34.0	60000.0	10%	NaN
H	Maketing	34.0	60000.0	10%	NaN
I	Maketing	26.0	100000.0	10%	NaN
H	Maketing	34.0	60000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN

ลบข้อมูล

- ลบทิ้งทั้งหมด
- ลบแถวบางส่วน
- ลบคอลัมน์บางส่วน
- ลบค่าซ้ำ

ลบทิ้งทั้งหมด

In [20]:
```
#นำเข้า DataFrame ใหม่
df = pd.read_csv("datasets/Employee.csv", index_col="Name")
df
```

Out[20]:

Name	Job	Age	Salary	Bonus	Address
A	Programmer	20.0	30000.0	10%	123.0
B	Programmer	18.0	NaN	10%	NaN
C	Developer	NaN	32000.0	10%	NaN
D	NaN	23.0	40000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN
F	Manager	NaN	75000.0	10%	NaN
NaN	NaN	NaN	NaN	NaN	NaN
H	Maketing	34.0	60000.0	10%	NaN
I	NaN	26.0	100000.0	10%	NaN
H	Maketing	34.0	60000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN

In [21]:
```
#ลบทิ้งทั้งหมด
df.dropna()
```

Out[21]:

Name	Job	Age	Salary	Bonus	Address
A	Programmer	20.0	30000.0	10%	123.0

ลบแถวบางส่วนที่มีค่าว่าง

In [22]:
```
#นำเข้า DataFrame ใหม่
df = pd.read_csv("datasets/Employee.csv", index_col="Name")
df
```

Out[22]:

Name	Job	Age	Salary	Bonus	Address
A	Programmer	20.0	30000.0	10%	123.0
B	Programmer	18.0	NaN	10%	NaN
C	Developer	NaN	32000.0	10%	NaN
D	NaN	23.0	40000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN
F	Manager	NaN	75000.0	10%	NaN
NaN	NaN	NaN	NaN	NaN	NaN
H	Maketing	34.0	60000.0	10%	NaN
I	NaN	26.0	100000.0	10%	NaN
H	Maketing	34.0	60000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN

In [23]:
```
#ลบแถวบางส่วนที่มีค่าว่าง
df.dropna(subset=['Age','Job'])
```

Out[23]:

Name	Job	Age	Salary	Bonus	Address
A	Programmer	20.0	30000.0	10%	123.0
B	Programmer	18.0	NaN	10%	NaN
E	Sale	29.0	40000.0	10%	NaN
H	Maketing	34.0	60000.0	10%	NaN
H	Maketing	34.0	60000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN

ลบคอลัมน์บางส่วนที่มีค่าว่าง

In [24]:
```
#นำเข้า DataFrame ใหม่
df = pd.read_csv("datasets/Employee.csv", index_col="Name")
df
```

Out[24]:

Name	Job	Age	Salary	Bonus	Address
A	Programmer	20.0	30000.0	10%	123.0
B	Programmer	18.0	NaN	10%	NaN
C	Developer	NaN	32000.0	10%	NaN
D	NaN	23.0	40000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN
F	Manager	NaN	75000.0	10%	NaN
NaN	NaN	NaN	NaN	NaN	NaN
H	Maketing	34.0	60000.0	10%	NaN
I	NaN	26.0	100000.0	10%	NaN
H	Maketing	34.0	60000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN

In [25]:
```
df.dropna(axis='columns')
```

Out[25]:

Name
A
B
C
D
E
F
NaN
H
I
H
E

ลบค่าซ้ำ

In [26]:
```
#นำเข้า DataFrame ใหม่
df = pd.read_csv("datasets/Employee.csv", index_col="Name")
df
```

Out[26]:

Name	Job	Age	Salary	Bonus	Address
A	Programmer	20.0	30000.0	10%	123.0
B	Programmer	18.0	NaN	10%	NaN
C	Developer	NaN	32000.0	10%	NaN
D	NaN	23.0	40000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN
F	Manager	NaN	75000.0	10%	NaN
NaN	NaN	NaN	NaN	NaN	NaN
H	Maketing	34.0	60000.0	10%	NaN
I	NaN	26.0	100000.0	10%	NaN
H	Maketing	34.0	60000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN

In [27]:
```
#เช็คค่าซ้ำ
df[df.duplicated]
```

Out[27]:

Name	Job	Age	Salary	Bonus	Address
H	Maketing	34.0	60000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN

In [28]:
```
#ลบค่าซ้ำ
df.drop_duplicates()
```

Out[28]:

Name	Job	Age	Salary	Bonus	Address
A	Programmer	20.0	30000.0	10%	123.0
B	Programmer	18.0	NaN	10%	NaN
C	Developer	NaN	32000.0	10%	NaN
D	NaN	23.0	40000.0	10%	NaN
E	Sale	29.0	40000.0	10%	NaN
F	Manager	NaN	75000.0	10%	NaN
NaN	NaN	NaN	NaN	NaN	NaN
H	Maketing	34.0	60000.0	10%	NaN
I	NaN	26.0	100000.0	10%	NaN

In [ ]: