

# Chapter 5 Data Preparing, Data Analytic and Data Visualization with Pandas

# Pandas คือ

- เป็นไลบรารีซอฟต์แวร์ที่เขียนขึ้นสำหรับภาษาโปรแกรม Python สำหรับการจัดการและวิเคราะห์ข้อมูล

# Data Frame คืออะไร?

Data frame คือ Excel table ประกอบด้วย

- column คือ ตัวแปรแต่ละตัว
- row คือ record เช่น ข้อมูลลูกค้าแต่ละคน หรือ transaction แต่ละอัน เป็นต้น
- ข้อมูลที่เก็บใน data frame ไม่จำเป็นต้องเป็นประเภทเดียวกัน

# Dataset คืออะไร

- ชุดข้อมูลที่ได้รวบรวมไว้ เพื่อนำมาวิเคราะห์ นำมาสอน (Train) ให้กับคอมพิวเตอร์เพื่อสร้างเป็น Model หรือใช้ทดสอบความถูกต้องแม่นยำของ Model

# แหล่งโหลด Dataset ฟรี

- <https://archive.ics.uci.edu/datasets>
- <https://www.kaggle.com/datasets>
- <https://datasetsearch.research.google.com/>

ฝึกโหลด dataset และดูรายละเอียด dataset


ไปที่เว็บ <https://www.kaggle.com/hesh97/titanicdataset-traincsv>

# โหลด dataset

Dataset


48


## Titanic-Dataset (train.csv)


 Syed Hamza Ali • updated 2 years ago (Version 1)

[Data](#) [Kernels \(16\)](#) [Discussion](#) [Activity](#) [Metadata](#)

Download (22 KB) [New Notebook](#)


 Usability 4.1

 License CC0: Public Domain

 Tags No tags yet

Data (22 KB)

Data Sources

 train.csv	891 x 12
---	----------

About this file

train.csv (Titanic-Dataset)

Columns

- # PassengerId
- # Survived
- # Pclass
- ▲ Name
- ▲ Sex
- ▲ Age
- # SibSp
- # Parch
- ▲ Ticket

train.csv (50.76 KB)

12 of 12 columns

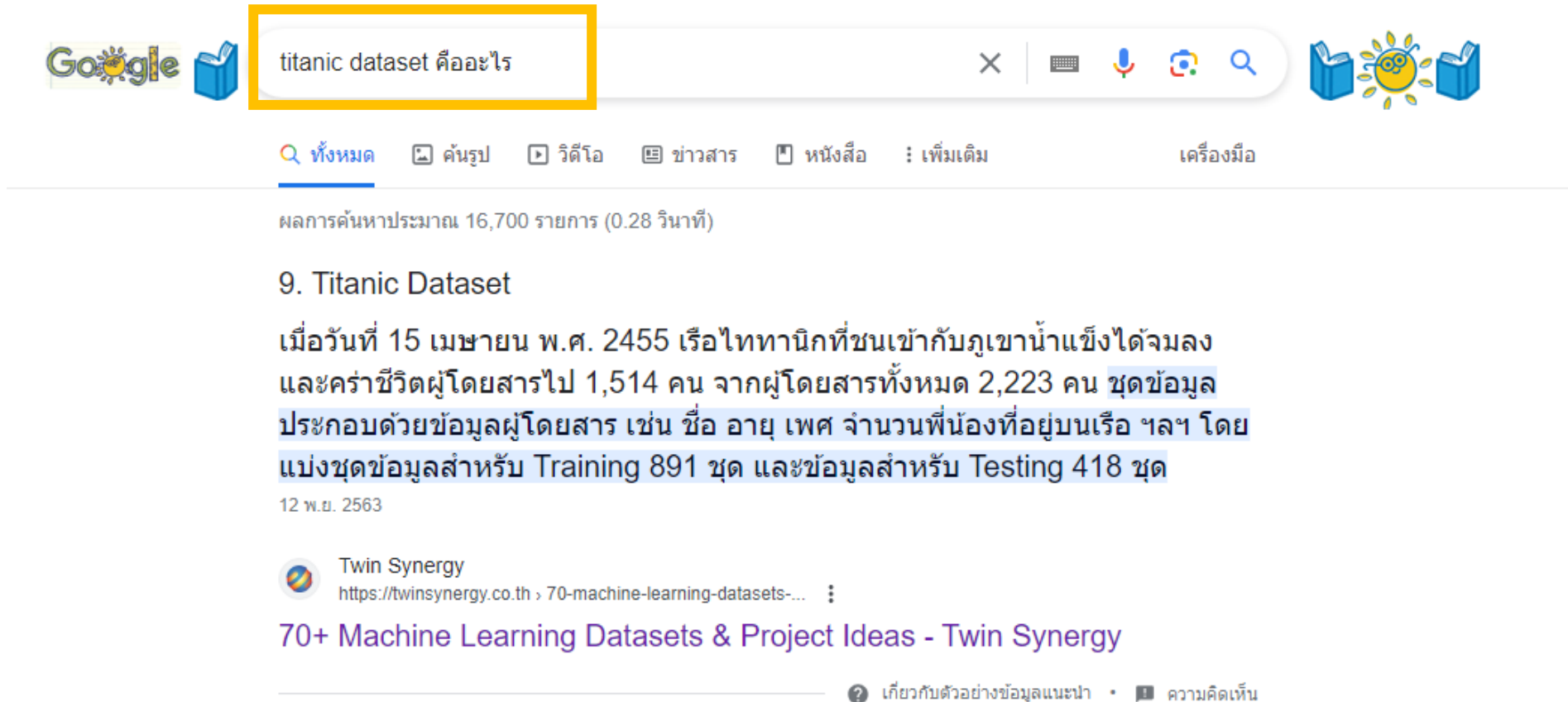
Views

# ทำความเข้าใจ dataset

	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mr.	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, Mrs.	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, Mrs.	female	26	0	0	STON/O2. 35101	7.925		S
5	4	1	1	Futrelle, Mrs.	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr.	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, Mr.	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy, Mr.	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, Mr.	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, Mrs.	female	27	0	2	347742	11.1333		S
11	10	1	2	Nasser, Mrs.	female	14	1	0	237736	30.0708		C
12	11	1	3	Sandstrom, Mrs.	female	4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, Mrs.	female	58	0	0	113783	26.55	C103	S
14	13	0	3	Saunders, Mr.	male	20	0	0	A/5. 2151	8.05		S
15	14	0	3	Andersson, Mr.	male	39	1	5	347082	31.275		S
16	15	0	3	Vestrom, Mrs.	female	14	0	0	350406	7.8542		S
17	16	1	2	Hewlett, Mrs.	female	55	0	0	248706	16		S
18	17	0	3	Rice, Master.	male	2	4	1	382652	29.125		Q
19	18	1	2	Williams, Mr.	male		0	0	244373	13		S
20	19	0	3	Vander Planck, Mrs.	female	31	1	0	345763	18		S
21	20	1	3	Masselman, Mrs.	female		0	0	2649	7.225		C
22	21	0	2	Fynney, Mr.	male	35	0	0	239865	26		S
23	22	1	2	Beesley, Mr.	male	34	0	0	248698	13	D56	S
24	23	1	3	McGowan, Mrs.	female	15	0	0	330923	8.0292		Q
25	24	1	1	Sloper, Mr.	male	28	0	0	113788	35.5	A6	S
26	25	0	3	Palsson, Mrs.	female	8	3	1	349909	21.075		S
27	26	1	3	Asplund, Mrs.	female	38	1	5	347077	31.3875		S
28	27	0	3	Emir, Mr.	male		0	0	2631	7.225		C
29	28	0	1	Fortune, Mr.	male	19	3	2	19950	263	C23 C25 C	S



# ค้นข้อมูลรายละเอียด dataset เพิ่มเติม



The image is a screenshot of a Google search results page. At the top, the Google logo is on the left, and a search bar in the center contains the text 'titanic dataset คืออะไร'. The search bar is highlighted with a yellow border. To the right of the search bar are icons for voice search, image search, and a magnifying glass. Below the search bar, there are tabs for 'ทั้งหมด' (All), 'ค้นรูป' (Image), 'วิดีโอ' (Video), 'ข่าวสาร' (News), 'หนังสือ' (Books), and 'เพิ่มเติม' (More). The 'ทั้งหมด' tab is selected. Below the tabs, the search results show 'ผลการค้นหาประมาณ 16,700 รายการ (0.28 วินาที)'. The first result is titled '9. Titanic Dataset' and contains a paragraph of Thai text: 'เมื่อวันที่ 15 เมษายน พ.ศ. 2455 เรือไททานิกที่ชนเข้ากับภูเขาน้ำแข็งได้จมลง และคร่าชีวิตผู้โดยสารไป 1,514 คน จากผู้โดยสารทั้งหมด 2,223 คน ชุดข้อมูล ประกอบด้วยข้อมูลผู้โดยสาร เช่น ชื่อ อายุ เพศ จำนวนพี่น้องที่อยู่บนเรือ ฯลฯ โดยแบ่งชุดข้อมูลสำหรับ Training 891 ชุด และข้อมูลสำหรับ Testing 418 ชุด'. Below the text is the date '12 พ.ย. 2563'. The result is from 'Twin Synergy' with the URL 'https://twinsynergy.co.th > 70-machine-learning-datasets-...'. The title of the result is '70+ Machine Learning Datasets & Project Ideas - Twin Synergy'. At the bottom, there is a footer with a question mark icon, the text 'เกี่ยวกับตัวอย่างข้อมูลเนาะนา', and a speech bubble icon with the text 'ความคิดเห็น'.

Google

titanic dataset คืออะไร

ทั้งหมด ค้นรูป วิดีโอ ข่าวสาร หนังสือ : เพิ่มเติม เครื่องมือ

ผลการค้นหาประมาณ 16,700 รายการ (0.28 วินาที)

## 9. Titanic Dataset

เมื่อวันที่ 15 เมษายน พ.ศ. 2455 เรือไททานิกที่ชนเข้ากับภูเขาน้ำแข็งได้จมลง และคร่าชีวิตผู้โดยสารไป 1,514 คน จากผู้โดยสารทั้งหมด 2,223 คน ชุดข้อมูล ประกอบด้วยข้อมูลผู้โดยสาร เช่น ชื่อ อายุ เพศ จำนวนพี่น้องที่อยู่บนเรือ ฯลฯ โดยแบ่งชุดข้อมูลสำหรับ Training 891 ชุด และข้อมูลสำหรับ Testing 418 ชุด

12 พ.ย. 2563

Twin Synergy  
https://twinsynergy.co.th > 70-machine-learning-datasets-...

### 70+ Machine Learning Datasets & Project Ideas - Twin Synergy

? เกี่ยวกับตัวอย่างข้อมูลเนาะนา • ความคิดเห็น

# ตัวอย่าง ผลการค้นเพิ่มเติมเกี่ยวกับ dataset

ชุดข้อมูลไททานิคประกอบด้วยคอลัมน์ต่างๆ ดังนี้

- Name - ชื่อเต็มของผู้โดยสาร
- Survived - เป็น 1 ถ้าผู้โดยสารรอดชีวิตจากเหตุโศกนาฏกรรมครั้งนี้; 0 ถ้าไม่รอด
- Pclass - ระดับชั้น เช่น 3 เป็นชั้นประหยัด; 1 เป็นชั้นหรูหรา
- Sex - เพศของผู้โดยสาร
- Age - อายุของผู้โดยสาร
- Sibsp - จำนวนพี่น้อง หรือสามีภรรยา ที่โดยสารมาด้วย
- Parch - จำนวนผู้ปกครอง หรือลูก ที่โดยสารมาด้วย
- Fare - ราคาตั๋ว
- Cabin - หมายเลขห้องโดยสาร
- Embarked - ท่าเรือที่ผู้โดยสารขึ้นมา
  - C = แซร์บอร์ก; Q = ควีนส์ทาวน์; S = เซาแทมปีตัน

# เตรียมข้อมูลและวิเคราะห์ข้อมูล ด้วย panda

## 1. import library

import pandas as pd

```
#1.import library  
import pandas as pd
```

## 2. import file

from google.colab import files

upload = files.upload()

```
#2.import file
```

```
from google.colab import files  
upload = files.upload()
```

Choose Files train.csv

- **train.csv**(text/csv) - 61194 bytes, last modified: 8/30/2023 - 100% done  
Saving train.csv to train.csv

### 3. อ่านไฟล์

```
titanic = pd.read_csv('train.csv')
```

```
#3. อ่านไฟล์  
titanic = pd.read_csv('train.csv')
```

[ Data Explore and Preparing ]

## 4. ข้อมูลที่โหลดมา

### titanic

#4. ข้อมูลที่โหลดมา  
titanic

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	sexConvert1	sexConvert2	
0	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2500	none	S	0	0
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38	1	0	PC 17599	71.2833	C85	C	1	1
2	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9250	none	S	1	1
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1000	C123	S	1	1
4	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.0500	none	S	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27	0	0	211536	13.0000	none	S	0	0
887	888	1	1	Graham, Miss. Margaret Edith	female	19	0	0	112053	30.0000	B42	S	1	1
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	30	1	2	W./C. 6607	23.4500	none	S	1	1
889	890	1	1	Behr, Mr. Karl Howell	male	26	0	0	111369	30.0000	C148	C	0	0
890	891	0	3	Dooley, Mr. Patrick	male	32	0	0	370376	7.7500	none	Q	0	0

891 rows × 14 columns

## 5. ตรวจสอบจำนวนแถว และคอลัมน์ของข้อมูล

titanic.shape

```
#5. ตรวจสอบจำนวนแถว และคอลัมน์ของข้อมูล  
titanic.shape
```

```
(891, 12)
```



## 6. แสดงข้อมูล 5 แถวแรก

titanic.head()

```
#6.แสดงข้อมูล 5 แถวแรก
```

```
titanic.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

## 7. กำหนดจำนวนแถว x ที่ต้องการให้ดึงออกมา ด้วย `.head(x)`

```
titanic.head(10) #7.สามารถกำหนดจำนวนแถวที่ให้ออกมาได้เช่นกัน
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C

## 8. คำสั่งที่ใช้เพื่อแสดงข้อมูล 5 แถวสุดท้าย ด้วย .tail()

titanic.tail() #9.เป็นคำสั่งที่ใช้เพื่อแสดงข้อมูล 5 แถวสุดท้าย

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
884	885	0	3	Sutehall, Mr. Henry Jr	male	25.0	0	0	SOTON/OQ 392076	7.050	NaN	S
885	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	382652	29.125	NaN	Q
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.450	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.750	NaN	Q

## 9. คำสั่งที่ใช้เพื่อแสดงข้อมูลแบบสุ่ม ด้วย `.sample()` :

```
titanic.sample(3) #8.เป็นคำสั่งที่ใช้เพื่อแสดงข้อมูล 3 แถวแบบสุ่ม
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
471	472	0	3	Cacic, Mr. Luka	male	38.0	0	0	315089	8.6625	NaN	S
284	285	0	1	Smith, Mr. Richard William	male	NaN	0	0	113056	26.0000	A19	S
823	824	1	3	Moor, Mrs. (Beila)	female	27.0	0	1	392096	12.4750	E121	S

## 10. ดูค่าทางสถิติของข้อมูล

titanic.describe()

```
#10. ดูค่าทางสถิติของข้อมูล  
titanic.describe()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

## 11. ตรวจสอบค่าว่าง

titanic.isnull().sum()

```
#11.ตรวจสอบค่าว่าง  
titanic.isnull().sum()
```

```
PassengerId      0  
Survived          0  
Pclass           0  
Name             0  
Sex              0  
Age            177  
SibSp            0  
Parch            0  
Ticket           0  
Fare             0  
Cabin           687  
Embarked         2  
dtype: int64
```

## 12. แทนที่ค่าว่างด้วยข้อมูลที่กำหนด ด้วยฟังก์ชัน `.fillna()`

```
#12.แทนที่ค่าว่างด้วยข้อมูลที่กำหนด ด้วยฟังก์ชัน fillna
#Age
titanic.Age.fillna('30', inplace =True )

#Cabin
titanic.Cabin.fillna('none', inplace = True )

#Embarked
titanic.Embarked.fillna('none', inplace = True )
```

เพิ่มเติม\* inplace=True ต่างกับ inplace=False อย่างไร

inplace=True จะแก้ไขข้อมูลใน Dataset เรา

inplace=False จะไม่แก้ไขข้อมูลใน Dataset เรา

ปกติถ้าเราไม่ใส่ inplace ในโค้ด จะมีค่าเป็น inplace=False

cr. <https://www.geeksforgeeks.org/what-does-inplace-mean-in-pandas/>



## 13. ตรวจสอบค่าว่างอีกครั้ง

```
#13. ตรวจสอบค่าว่างอีกครั้ง  
titanic.isnull().sum()
```

```
PassengerId    0  
Survived        0  
Pclass          0  
Name            0  
Sex             0  
Age            0  
SibSp           0  
Parch           0  
Ticket         0  
Fare           0  
Cabin          0  
Embarked        0  
dtype: int64
```

## info() : แสดงข้อมูลในภาพรวม

- เป็นคำสั่งใช้แสดงข้อมูลในภาพรวมที่สำคัญเกี่ยวกับ data ที่เราสนใจ เช่น จำนวนแถวและคอลัมน์ของข้อมูล, จำนวนข้อมูลที่ไม่เป็นค่าว่าง (Null), ประเภทของข้อมูลที่เก็บอยู่ในแต่ละคอลัมน์ หรือ memory ที่ใช้
- หากเราเข้าใจข้อมูลเบื้องต้นแล้ว จะสามารถจัดการกับ data ได้ง่ายขึ้น
- เช่น ข้อมูลบาง column เป็นตัวเลข แต่อาจถูกเก็บเป็น string จึงต้องถูกแปลงก่อนที่จะเอามาวงกลมบวกลบคูณหารได้

## 14. info() : แสดงข้อมูลในภาพรวม

#14. แสดงข้อมูลในภาพรวม

`titanic.info()`

# เป็นคำสั่งใช้แสดงข้อมูลในภาพรวมที่สำคัญเกี่ยวกับ data ที่เราสนใจ

# เช่น จำนวนแถวและคอลัมน์ของข้อมูล, จำนวนข้อมูลที่ไม่เป็นค่าว่าง (Null), ประเภทของข้อมูลที่เก็บอยู่ในแต่ละคอลัมน์

# หรือ memory ที่ใช้

# หากเราเข้าใจข้อมูลเบื้องต้นแล้ว จะสามารถจัดการกับ data ได้ง่ายขึ้น

# เช่น ข้อมูลบาง column เป็นตัวเลข แต่อาจถูกเก็บเป็น string จึงต้องถูกแปลงก่อนที่จะเอามาวกอบคุณหาได้

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 891 entries, 0 to 890
```

```
Data columns (total 12 columns):
```

#	Column	Non-Null Count	Dtype
0	PassengerId	891 non-null	int64
1	Survived	891 non-null	int64
2	Pclass	891 non-null	int64
3	Name	891 non-null	object
4	Sex	891 non-null	object
5	Age	891 non-null	object
6	SibSp	891 non-null	int64
7	Parch	891 non-null	int64
8	Ticket	891 non-null	object
9	Fare	891 non-null	float64
10	Cabin	891 non-null	object
11	Embarked	891 non-null	object

```
dtypes: float64(1), int64(5), object(6)
```

```
memory usage: 83.7+ KB
```

# Panda Data Types

## Pandas Data Types

Pandas dtype	Python type	NumPy type
object	str or mixed	string_, unicode_, mixed types
int64	int	int_, int8, int16, int32, int64, uint8, uint16, uint32, uint64
float64	float	float_, float16, float32, float64
bool	bool	bool_

## 15. เปลี่ยนชนิดข้อมูล

```
titanic['Age'] = titanic['Age'].astype('int')
```

```
#15. เปลี่ยนชนิดข้อมูล  
titanic['Age'] = titanic['Age'].astype('int')
```

## 16. ตรวจสอบข้อมูล หลังเปลี่ยน

titanic.info()

```
#16.ตรวจสอบข้อมูล หลังเปลี่ยน  
titanic.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 891 entries, 0 to 890  
Data columns (total 12 columns):  
#   Column      Non-Null Count  Dtype  
---  ---  
0   PassengerId 891 non-null   int64  
1   Survived    891 non-null   int64  
2   Pclass      891 non-null   int64  
3   Name        891 non-null   object  
4   Sex         891 non-null   object  
5   Age         891 non-null   int64  
6   SibSp       891 non-null   int64  
7   Parch       891 non-null   int64  
8   Ticket      891 non-null   object  
9   Fare        891 non-null   float64  
10  Cabin       891 non-null   object  
11  Embarked    891 non-null   object  
dtypes: float64(1), int64(6), object(5)  
memory usage: 83.7+ KB
```

17. แสดงรายละเอียดว่า dataset มี column อะไรบ้าง ด้วย `columns`

```
#17.แสดงรายละเอียดว่ามี column อะไรอยู่บ้างใน data  
titanic.columns
```

```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',  
      'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],  
      dtype='object')
```

## 18. เลือกข้อมูลบาง Column

#18.เลือกข้อมูลบาง Column

```
a = titanic[['Embarked', 'Fare']]
```

a

	Embarked	Fare
0	S	7.2500
1	C	71.2833
2	S	7.9250
3	S	53.1000
4	S	8.0500
...	...	...
886	S	13.0000
887	S	30.0000
888	S	23.4500
889	C	30.0000
890	Q	7.7500

891 rows × 2 columns



[ Data Analytic ]

## 19. เรียงข้อมูล โดยเรียงตามคอลัมน์ที่ระบุ

`titanic.sort_values(by=['Age'], ascending=True , inplace=False)`

#19. เรียงข้อมูล

```
titanic.sort_values(by=['Age'], ascending=True , inplace=False)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
644	645	1	3	Baclini, Miss. Eugenie	female	0	2	1	2666	19.2583	none	C
78	79	1	2	Caldwell, Master. Alden Gates	male	0	0	2	248738	29.0000	none	S
469	470	1	3	Baclini, Miss. Helene Barbara	female	0	2	1	2666	19.2583	none	C
831	832	1	2	Richards, Master. George Sibley	male	0	1	1	29106	18.7500	none	S
305	306	1	1	Allison, Master. Hudson Trevor	male	0	1	2	113781	151.5500	C22 C26	S
...	...	...	...	...	...	...	...	...	...	...	...	...
745	746	0	1	Crosby, Capt. Edward Gifford	male	70	1	1	WE/P 5735	71.0000	B22	S
493	494	0	1	Artagaveytia, Mr. Ramon	male	71	0	0	PC 17609	49.5042	none	C
96	97	0	1	Goldschmidt, Mr. George B	male	71	0	0	PC 17754	34.6542	A5	C
851	852	0	3	Svensson, Mr. Johan	male	74	0	0	347060	7.7750	none	S
630	631	1	1	Barkworth, Mr. Algernon Henry Wilson	male	80	0	0	27042	30.0000	A23	S

891 rows × 12 columns

## 20. การหาค่าของข้อมูลกลุ่มย่อยโดยใช้ Groupby

titanic.groupby(by='Pclass').min()

```
#20.การหาค่าของข้อมูลกลุ่มย่อยโดยใช้ Groupby  
titanic.groupby(by='Pclass').min()
```

	PassengerId	Survived	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
Pclass											
1	2	0	Allen, Miss. Elisabeth Walton	female	0	0	0	110152	0.0	A10	C
2	10	0	Abelson, Mr. Samuel	female	0	0	0	11668	0.0	D	C
3	1	0	Abbing, Mr. Anthony	female	0	0	0	12460	0.0	E10	C

## 21. ถ้าต้องการแสดงผลหลายค่า ใช้คำสั่ง .agg()

titanic.groupby(by='Pclass').agg(['mean', 'min', 'max'])

```
#21.ถ้าต้องการแสดงผลหลายค่า ใช้คำสั่ง .agg
```

```
titanic.groupby(by='Pclass').agg(['mean', 'min', 'max'])
```

```
<ipython-input-26-1e7140517681>:2: FutureWarning: ['Name', 'Sex', 'Ticket', 'Cabin', 'Embarked'] did not aggregate successfully. If  
titanic.groupby(by='Pclass').agg(['mean', 'min', 'max'])
```

	PassengerId			Survived			Age		SibSp			Parch			Fare			
	mean	min	max	mean	min	max	mean	min	max	mean	min	max	mean	min	max	mean	min	max
Pclass																		
1	461.597222	2	890	0.629630	0	1	37.083333	0	80	0.416667	0	3	0.356481	0	4	84.154687	0.0	512.3292
2	445.956522	10	887	0.472826	0	1	29.864130	0	70	0.402174	0	3	0.380435	0	3	20.662183	0.0	73.5000
3	439.154786	1	891	0.242363	0	1	26.468432	0	74	0.615071	0	8	0.393075	0	6	13.675550	0.0	69.5500

## 22. เลือกจำนวนแถวที่ต้องการ ด้วย `iloc[]`

`titanic.groupby(by='Pclass').agg(['mean', 'min', 'max']).iloc[0:2]`

#22. เลือกจำนวนแถวที่ต้องการ

```
titanic.groupby(by='Pclass').agg(['mean', 'min', 'max']).iloc[0:2]
```

<ipython-input-27-5f0635ff1163>:2: FutureWarning: ['Name', 'Sex', 'Ticket', 'Cabin', 'Embarked'] did not aggregate successfully. If  
titanic.groupby(by='Pclass').agg(['mean', 'min', 'max']).iloc[0:2]

	PassengerId			Survived			Age			SibSp			Parch			Fare		
	mean	min	max	mean	min	max	mean	min	max	mean	min	max	mean	min	max	mean	min	max
Pclass																		
1	461.597222	2	890	0.629630	0	1	37.083333	0	80	0.416667	0	3	0.356481	0	4	84.154687	0.0	512.3292
2	445.956522	10	887	0.472826	0	1	29.864130	0	70	0.402174	0	3	0.380435	0	3	20.662183	0.0	73.5000

## 23. เลือกคอลัมน์ที่ต้องการ ด้วย [col\_list].iloc[]

```
col_list = ['Parch', 'Fare']
```

```
titanic.groupby(by='Pclass').agg(['mean', 'min', 'max'])[col_list].iloc[0:5]
```

#23. เลือกคอลัมน์ที่ต้องการ

```
col_list = ['Parch', 'Fare']
```

```
titanic.groupby(by='Pclass').agg(['mean', 'min', 'max'])[col_list].iloc[0:5]
```

```
<ipython-input-30-c0bfe4a28644>:3: FutureWarning: ['Name', 'Sex', 'Ticket', 'Cabin', 'Embarked'] did not aggregate
titanic.groupby(by='Pclass').agg(['mean', 'min', 'max'])[col_list].iloc[0:5]
```

	Parch			Fare		
	mean	min	max	mean	min	max
Pclass						
1	0.356481	0	4	84.154687	0.0	512.3292
2	0.380435	0	3	20.662183	0.0	73.5000
3	0.393075	0	6	13.675550	0.0	69.5500

## 24. การเลือกข้อมูลด้วย Condition

titanic[titanic['Pclass'] > 2 ]

```
#24.การเลือกข้อมูลด้วย Condition  
titanic[titanic['Embarked'] == 'Q']
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
5	6	0	3	Moran, Mr. James	male	30	0	0	330877	8.4583	none	Q
16	17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.1250	none	Q
22	23	1	3	McGowan, Miss. Anna "Annie"	female	15	0	0	330923	8.0292	none	Q
28	29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female	30	0	0	330959	7.8792	none	Q
32	33	1	3	Glynn, Miss. Mary Agatha	female	30	0	0	335677	7.7500	none	Q
...	...	...	...	...	...	...	...	...	...	...	...	...
790	791	0	3	Keane, Mr. Andrew "Andy"	male	30	0	0	12460	7.7500	none	Q
825	826	0	3	Flynn, Mr. John	male	30	0	0	368323	6.9500	none	Q
828	829	1	3	McCormack, Mr. Thomas Joseph	male	30	0	0	367228	7.7500	none	Q
885	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39	0	5	382652	29.1250	none	Q
890	891	0	3	Dooley, Mr. Patrick	male	32	0	0	370376	7.7500	none	Q

77 rows × 12 columns

## 25. กรณี > 1 Condition

`titanic[(titanic['Survived'] == 1) | (titanic['Pclass'] == 3)]`

#25.กรณี > 1 Condition

```
titanic[(titanic['Survived'] == 1) | (titanic['Pclass'] == 3)]
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
885	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	382652	29.1250	NaN	Q
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

714 rows × 12 columns



## 26. การใช้ Function ร่วมกับ Dataframe

#step1 : สร้างฟังก์ชัน

```
def rating_function(x):
```

```
    if x == 'male':
```

```
        return "0"
```

```
    else:
```

```
        return "1"
```

#step2 : เรียกใช้ฟังก์ชัน

```
titanic['sexConvert1'] = titanic['Sex'].apply(rating_function)
```

```
titanic.head()
```

#26.การใช้ Function ร่วมกับ Dataframe

#สร้างฟังก์ชัน

```
def rating_function(x):  
    if x == 'male':  
        return "0"  
    else:  
        return "1"
```

#เรียกใช้ฟังก์ชัน

```
titanic['sexConvert1'] = titanic['Sex'].apply(rating_function)  
titanic.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	sexConvert1
0	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2500	none	S	0
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38	1	0	PC 17599	71.2833	C85	C	1
2	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9250	none	S	1
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1000	C123	S	1
4	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.0500	none	S	0

## 27. การใช้ lambda Function ร่วมกับ Dataframe

```
titanic['sexConvert2'] = titanic['Sex'].apply(lambda x: '0' if x == 'male' else '1')
```

```
titanic.head()
```

#27.การใช้ lambda Function ร่วมกับ Dataframe

```
titanic['sexConvert2'] = titanic['Sex'].apply(lambda x: '0' if x == 'male' else '1')
```

```
titanic.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	sexConvert1	sexConvert2
0	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.2500	none	S	0	0
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38	1	0	PC 17599	71.2833	C85	C	1	1
2	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.9250	none	S	1	1
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1000	C123	S	1	1
4	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.0500	none	S	0	0

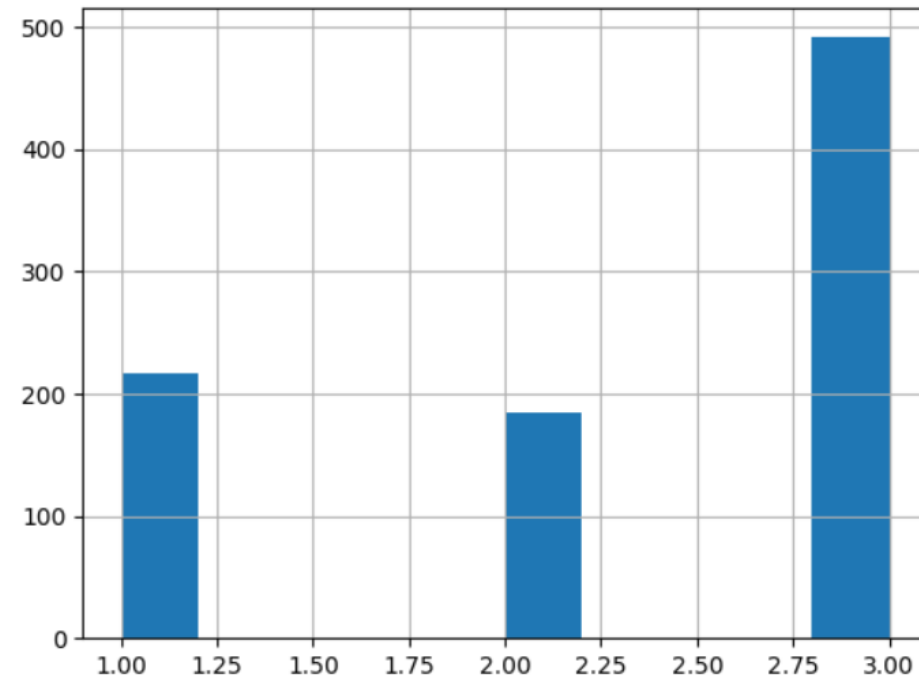
[ Data Visualization ]

## 28. Plot Histogram ด้วย hist()

titanic['Pclass'].hist()

```
#28.Plot Histogram ด้วย hist()  
#ex1  
titanic['Pclass'].hist()
```

<Axes: >

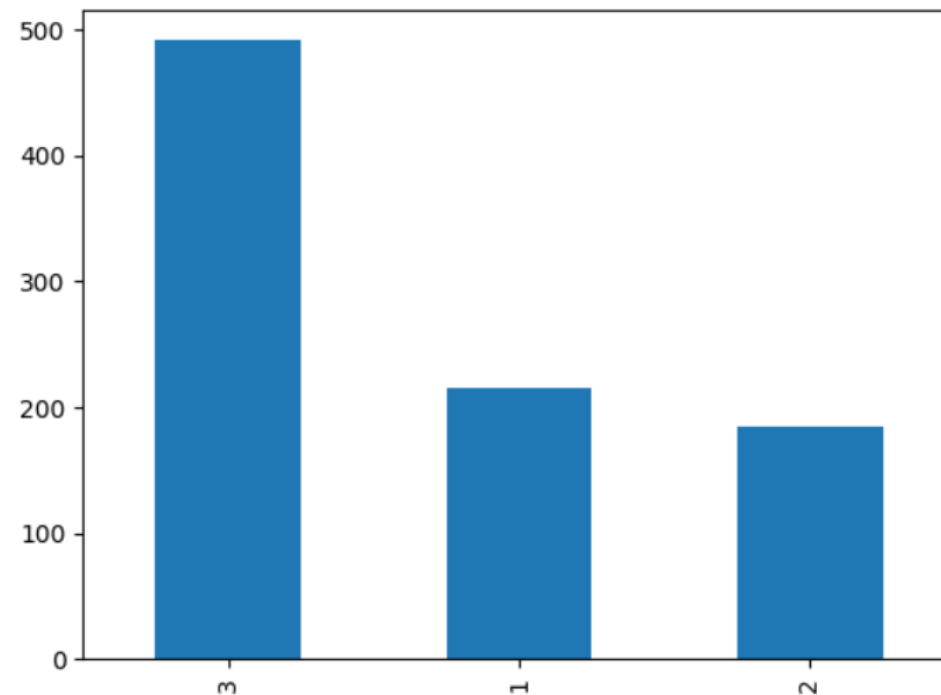


## 29. Plot Bar Charts ด้วย value\_counts().plot.bar()

titanic['Pclass'].value\_counts().plot.bar()

```
#29. Plot Bar Charts ด้วย value_counts().plot.bar()
#ex1
titanic['Pclass'].value_counts().plot.bar()
```

<Axes: >

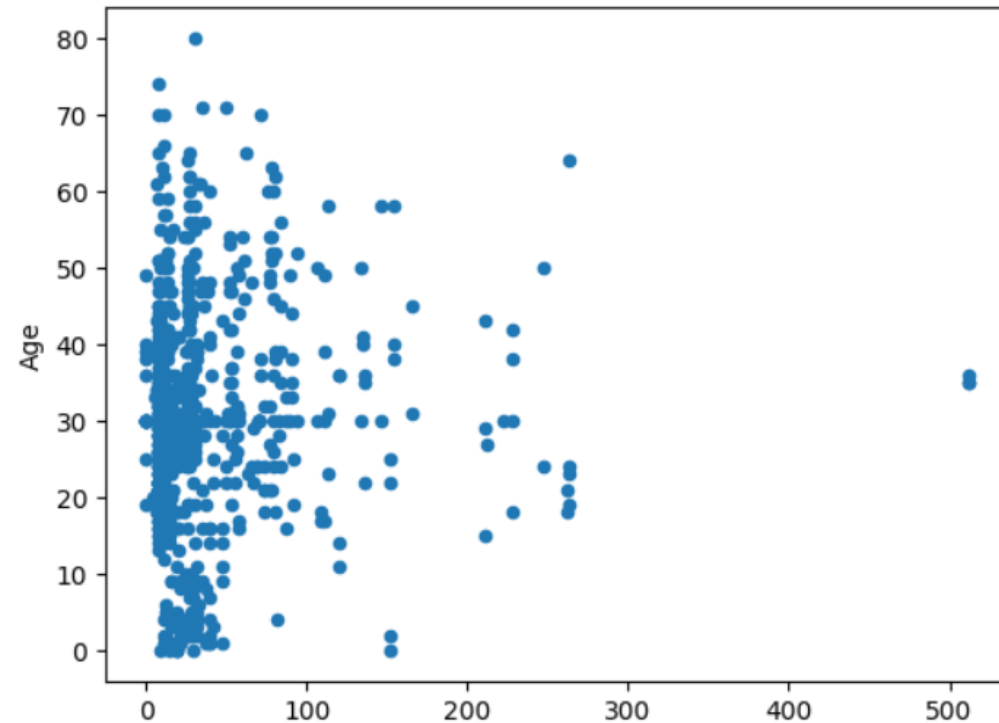


### 30. plot scatter plot ดูการกระจายข้อมูล และค่าผิดปกติ (outlier)

titanic.plot.scatter(x='Fare', y='Age')

```
#31.plot scatter plot ดูการกระจายข้อมูล และค่าผิดปกติ (outlier)  
titanic.plot.scatter(x='Fare', y='Age')
```

<Axes: xlabel='Fare', ylabel='Age'>



# แหล่งศึกษาด้วยตัวเองเพิ่มเติม

- ค้น google >> Pandas cheat sheet
- <https://pandas.pydata.org/>



# ใบงาน

- ให้แบ่งกลุ่มๆ ละ 7 คน (4 กลุ่ม)
- ให้แต่ละกลุ่มเลือก dataset มา 1 dataset ห้ามซ้ำกัน โดยค้นจากเว็บ

<https://archive.ics.uci.edu/datasets>

<https://www.kaggle.com/datasets>

<https://datasetsearch.research.google.com/> หรือเว็บอื่น ๆ

- ทำความเข้าใจ dataset ว่าประกอบด้วยคอลัมภ์อะไรบ้าง แต่ละคอลัมภ์คือข้อมูลอะไร และวิเคราะห์ข้อมูลใน dataset ที่ได้ (เลือกใช้คำสั่งจากเนื้อหา pandas ที่ได้เรียนจำนวน 14 คำสั่ง : คนละ 2 คำสั่ง) และอธิบายว่าสามารถนำข้อมูลที่วิเคราะห์ได้ในแต่ละคำสั่งนั้นไปใช้ประโยชน์อย่างไร
- มานำเสนอสัปดาห์หน้า นำเสนอทุกคนในกลุ่ม ตามที่ตนเองรับผิดชอบคำสั่ง

# Reference

Botnoi. (2020). **Data Science Essential Course**. Bangkok : BotNoi.

Chris Moffitt . (2018). **Overview of Pandas Data Types**. Site on

[https://pbpython.com/pandas\\_dtypes.html](https://pbpython.com/pandas_dtypes.html)

Dpsvnshob130491. (2023). **What does inplace mean in Pandas?** Site on

<https://www.geeksforgeeks.org/what-does-inplace-mean-in-pandas/>