

Python 0 到 1 基礎商業分析實戰

TMR

Pandas教學

大綱

1. `pd.DataFrame()`
2. `head()`
3. `info()`
4. `shape`
5. `df.columns` 與命名方式
6. `df[‘變數’]`

1. `mean()` 、 `max()` 、 `sum()`
2. `describe()`
3. `sort_values()`
4. `df.drop`
5. `df.values.tolist()`
6. `df.astype`

Dataframe

1. 一個兩個維度的資料結構，它是一個自帶擁有大量常見資料分析。連分布式運算架構PySpark 也是以Dataframe為基礎架構
2. 使用表格思維！方便我們快速操作數據
3. 基本上只要把它當成指令操作的Excel 就可以了

想像一下

Columns

rows

Regd. No	Name	Marks%
1000	Steve	86.29
1001	Mathew	91.63
1002	Jose	72.90
1003	Patty	69.23
1004	Vin	88.30

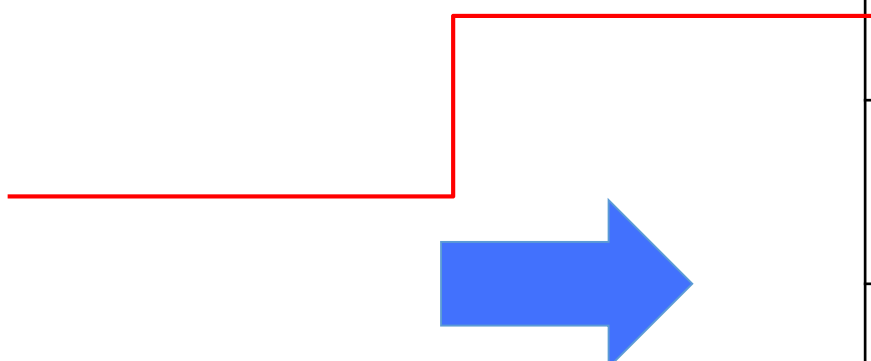
自建資料 - member

member : 會員資料

1. uid : 使用者ID
2. age : 年齡
3. name : 姓名

pd.DataFrame()

```
{  
'A' : [1, 1, 4],  
'B' : [2, 2, 5],  
'C' : [3, 3, 3]  
}
```



A	B	C
1	2	3
1	2	3
4	5	3

常見顯示資訊

- 顯示前五筆資料：head()
- 顯示欄位資訊：info()

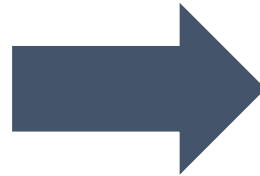
Df.shape

- 列出共有幾欄幾列
- (列數,欄數) (R,C) 直欄橫列

修改欄位名稱

A	B	C
1	2	3
1	2	3
4	5	3

df.columns =
['col1','col2','col3']

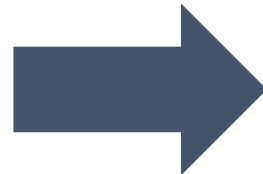


col1	col2	col3
1	2	3
1	2	3
4	5	3

取值 `df['colume']`

A	B	C
1	2	3
1	2	3
4	5	3

`df['B']`



B
2
2
5

常見計算

1. 整欄資料統計資訊：`df['欄位名稱'].describe()`
2. 整欄平均：`df['欄位名稱'].mean()`
3. 整欄最大值：`df['欄位名稱'].max()`
4. 整欄加總：`df['欄位名稱'].sum()`

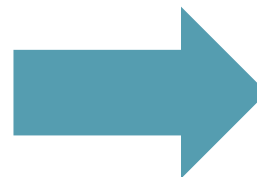
sort_values('colume')

根據某個欄位，進行資料排序。

```
df.sort_values('freq',ascending = False)
```

RF_table - DataFrame

Index	0-7 da
1 freq	4
2 freq	12
3 freq	15
4 freq	16
5 freq	16
>5 freq	13

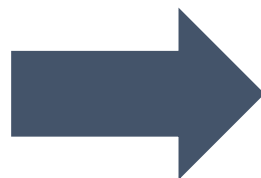


Index	
>5 freq	>5 freq
5 freq	5 freq
4 freq	4 freq
3 freq	3 freq
2 freq	2 freq
1 freq	1 freq

移除欄位 drop

A	B	C
1	2	3
1	2	3
4	5	3

df.drop (columns=['B'])



A	C
1	3
1	3
4	3

強制型轉

- `Df.values.tolist()`
- `Df['column'].astype('float64')`
- `Df['column'].astype('int')`