빅데이터 실습

7주차 1차시

Pandas Cheat Sheet 한장으로 Pandas 복습하기

Pandas Cheat Sheet 한장으로 Pandas 복습하기

학습개요

1/ Pandas Cheat Sheet 소개

2/ Pandas Cheat Sheet로 복습하기

Pandas Cheat Sheet 한장으로 Pandas 복습하기

학습목표

1/ Pandas Cheat Sheet가 무엇인지 설명할 수 있다.

2/ Pandas의 기본 문법들을 설명할 수 있다.

01

Pandas Cheat Sheet 소개

1 Pandas Cheat Sheet 소개

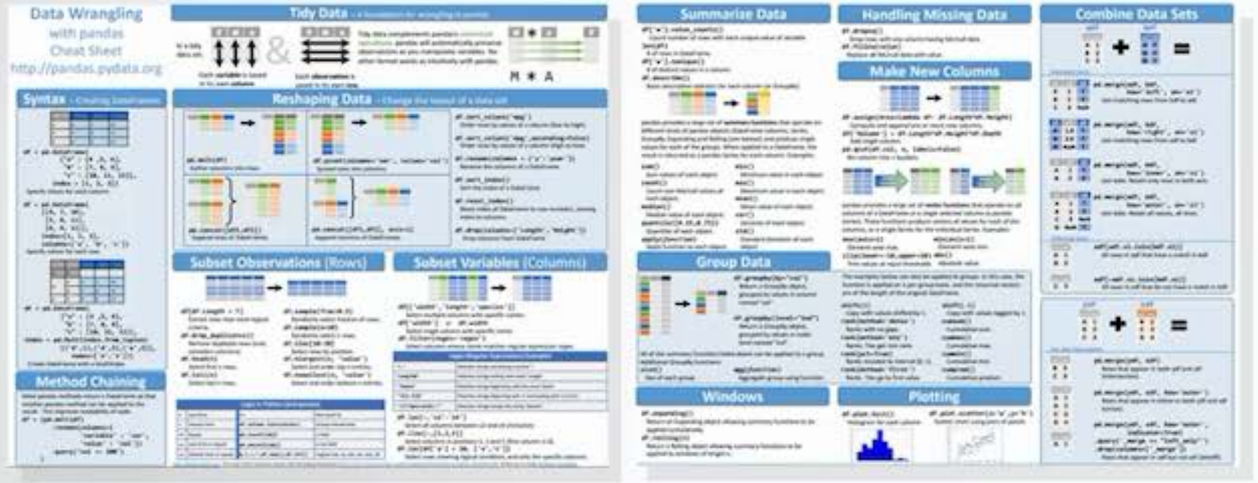
Pandas Cheat Sheet란?

- ✓ Pandas에서 공식적으로 제공하는 커닝페이퍼
- ✓ Pandas 문법 중 가장 많이 활용되는 주요 문법들을 단, 2장으로 요약해 놓은 문서

① Pandas Cheat Sheet 소개

Pandas Cheat Sheet란?

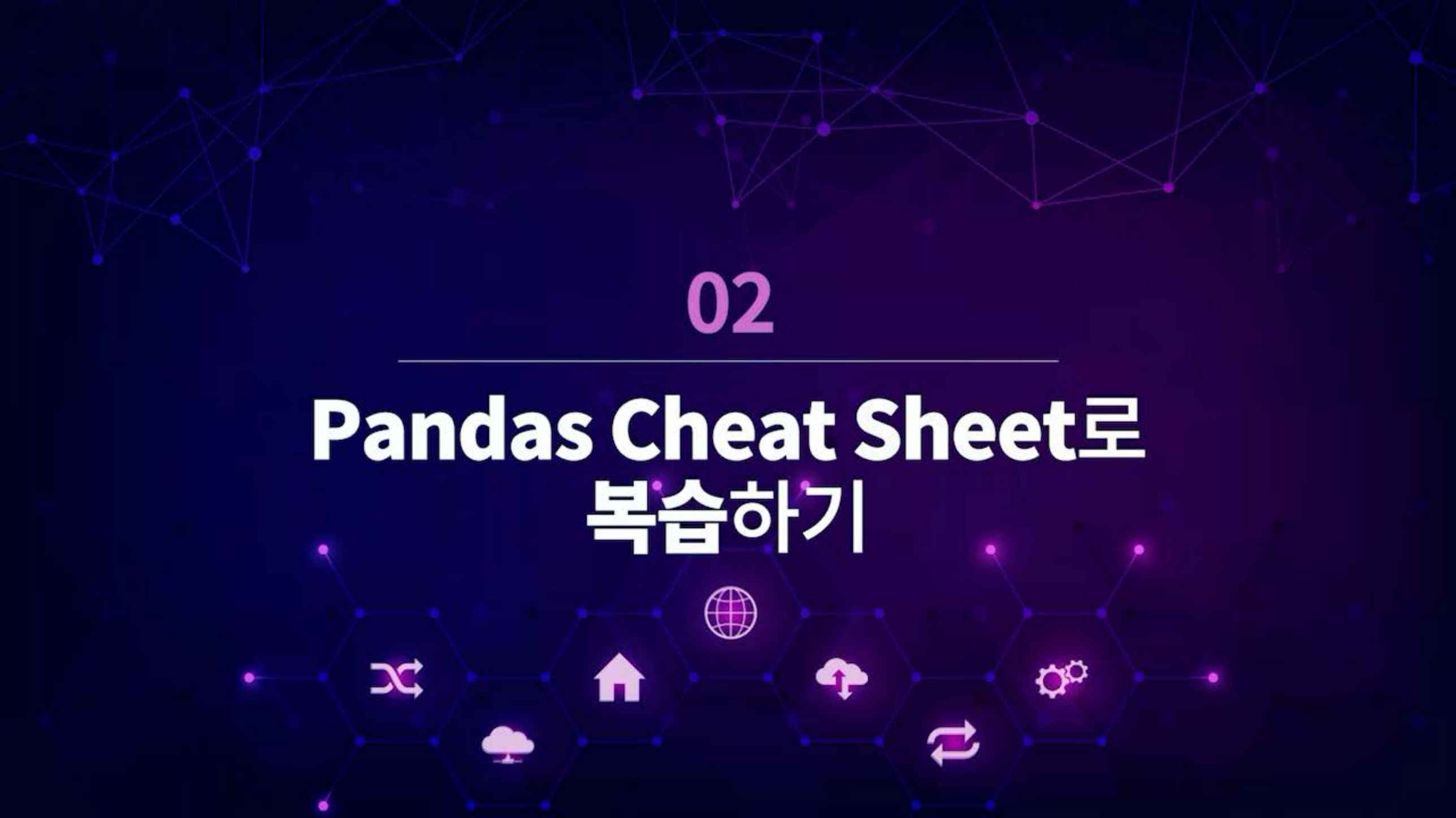
✓ 판다스의 공식 홈페이지에서 다운로드 가능



〈출처〉 https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf

02

Pandas Cheat Sheet로 복습하기



② Pandas Cheat Sheet로 복습하기

7 DataFrame 생성하기

✓ Syntax - Creating DataFrames

	a	b	c
1	4	7	10
2	5	8	11
3	6	9	12

```
df = pd.DataFrame(
        {"a" : [4, 5, 6],
         "b" : [7, 8, 9],
         "c" : [10, 11, 12]},
        index = [1, 2, 3])
```

② Pandas Cheat Sheet로 복습하기

7 DataFrame 생성하기

✓ Syntax - Creating DataFrames

```
df = pd.DataFrame(
        {"a" : [4, 5, 6],
         "b" : [7, 8, 9],
         "c" : [10, 11, 12]},
index = pd.MultiIndex.from_tuples(
        [('d',1),('d',2),('e',2)],
         names=['n','v']))
```

		a	b	c
n	v			
d	1	4	7	10
	2	5	8	11
e	3	6	9	12

② Pandas Cheat Sheet로 복습하기

7 DataFrame 생성하기

✓ Syntax - Creating DataFrames

- ◉ 리스트 형태의 데이터 타입은 모두 그 데이터의 인자로 활용 가능
- ◉ index와 columns 인자 같은 경우에는 지정하지 않으면 0부터 시작해서 1씩 증가하는 값으로 설정
- ◉ DataFrame을 생성하는 것보다 read_excel이나 read_csv와 같이 외부에 있는 파일을 크롤링하는 경우를 빈번하게 사용

② Pandas Cheat Sheet로 복습하기

데이터 모양 바꾸기 (Reshaping Data)

✓ pd.melt(df)

```
df
```

	A	B	C
0	a	1	2
1	b	3	4
2	c	5	6

```
#df.melt()
pd.melt(df)
```

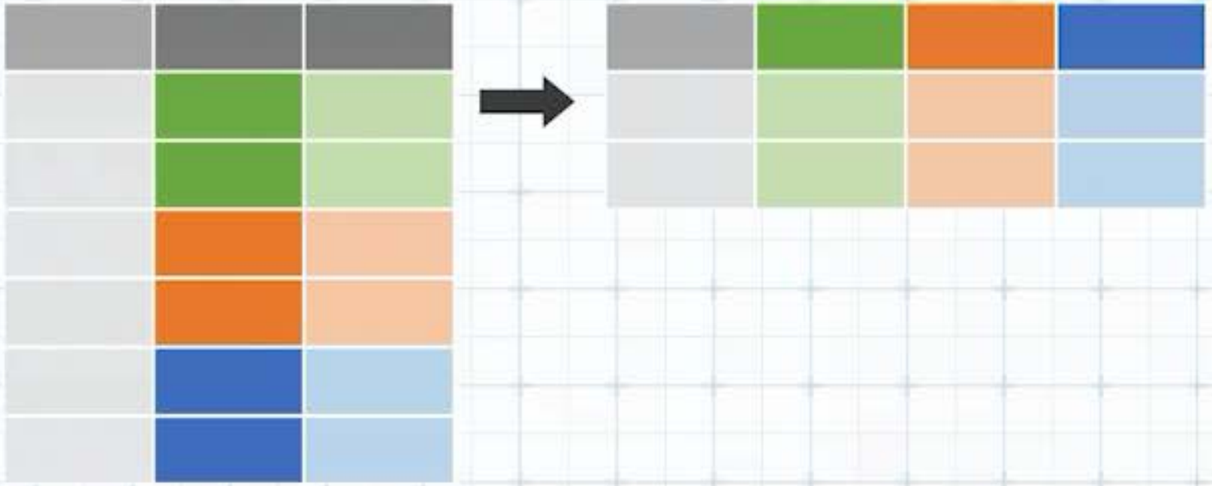
	variable	value
0	A	a
1	A	b
2	A	c
3	B	1
4	B	3
5	B	5
6	C	2
7	C	4
8	C	6

② Pandas Cheat Sheet로 복습하기

데이터 모양 바꾸기 (Reshaping Data)

✓ df.pivot(columns='var', values='val')

◉ 데이터 모양 바꾸는 함수



pivot_table

그룹집계

pivot

데이터의 모양만 바꿔줌

② Pandas Cheat Sheet로 복습하기

7 데이터 모양 바꾸기 (Reshaping Data)

✓ df.pivot(columns='var', values='val')

df

	foo	bar	baz	zoo
0	one	A	1	x
1	one	B	2	y
2	one	C	3	z
3	two	A	4	q
4	two	B	5	w
5	two	C	6	t

```
df.pivot(index='foo', columns='bar', values='baz')
```

bar	A	B	C
foo			
one	1	2	3
two	4	5	6

◉ row에 있는 것들을 column으로 변경해 주는 함수

② Pandas Cheat Sheet로 복습하기

데이터 모양 바꾸기 (Reshaping Data)

✓ df.pivot(columns='var', values='val')

```
df.pivot(index='foo', columns='bar', values='baz')
```

bar	A	B	C
foo			
one	1	2	3
two	4	5	6

동일한 결과

```
df.set_index(['foo','bar'])['baz'].unstack()
```

bar	A	B	C
foo			
one	1	2	3
two	4	5	6

여러 가지 구문들을 잘 활용할 수 있는 능력을 키우시는 게 상당히 중요하다고 할 수 있습니다.

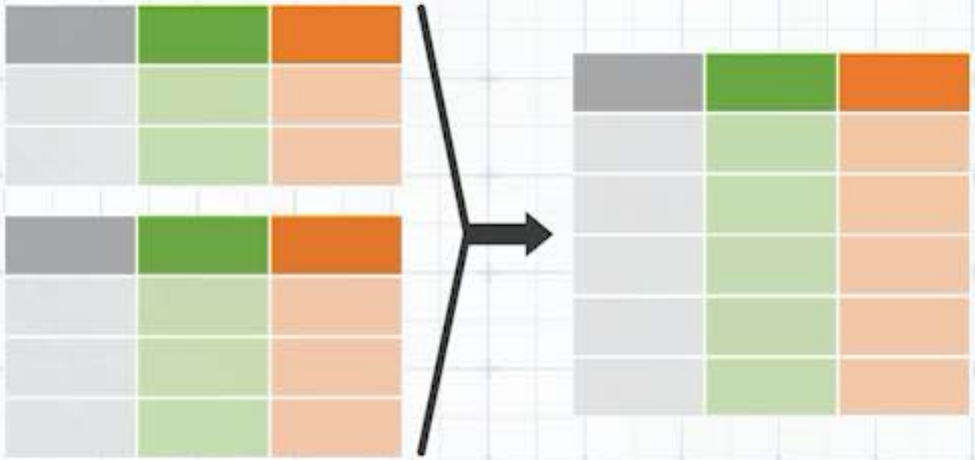
② Pandas Cheat Sheet로 복습하기

데이터 모양 바꾸기 (Reshaping Data)

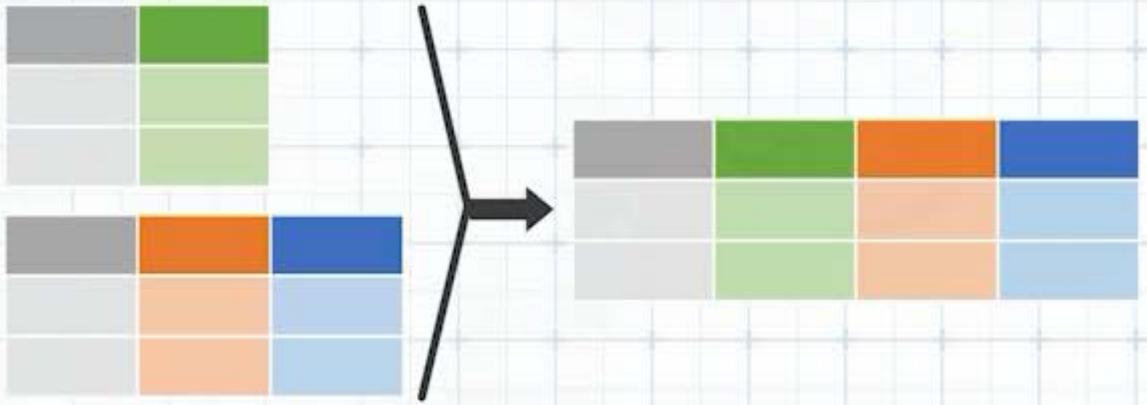
✓ pd.concat([df1,df2], axis=0(1))

◉ 2개의 데이터프레임을 이어 붙일 때(concatenate)사용하는 함수

row 방향으로 합칠 때



columns 방향으로 합칠 때



② Pandas Cheat Sheet로 복습하기

데이터 모양 바꾸기 (Reshaping Data)

✓ df.sort_values('mpg') mpg 기준 컬럼

◉ 값을 정렬하는 함수

✓ df.sort_index( )

◉ 인덱스를 정렬하는 함수

✓ df.rename(columns = {'y':'year'})

◉ columns의 이름을 변경해주는 함수

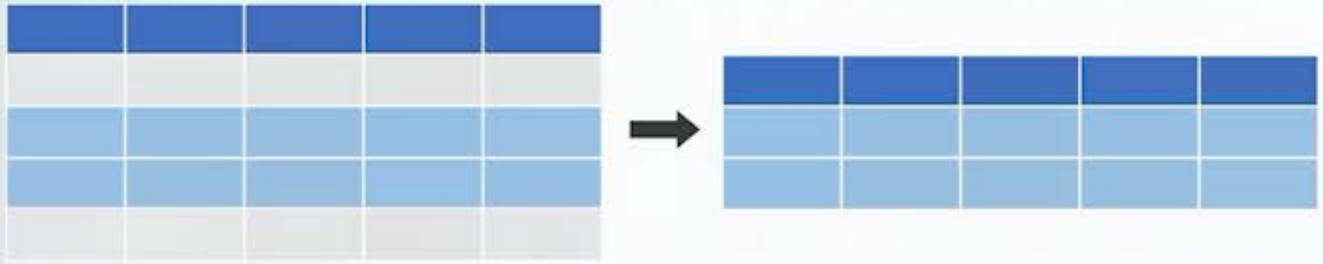
✓ df.drop(columns=['Length','Height'])

◉ columns을 삭제하는 함수 (row 단위로도 가능)

② Pandas Cheat Sheet로 복습하기

ㄱ 색인 기능 (Subset Observations(Rows))

✓ 전체 데이터프레임에서
원하는 데이터프레임만 선택하는 방식



② Pandas Cheat Sheet로 복습하기

색인 기능 (Subset Observations(Rows))

✓ df.drop_duplicates( )

◉ 특정 columns 값들 중에서
중복되어 있는 것들을 제거하고 값을 리턴하는 함수

예

100명의 학생들이 1, 2, 3반으로 나누어져 있을 때

1, 2, 3을 리턴함

② Pandas Cheat Sheet로 복습하기

색인 기능 (Subset Observations(Rows))

✓ df.sample(frac=0.5)

◉ 전체 데이터프레임에서
랜덤으로 몇 개를 추출해 주는 함수

◉ frac

➡ 비율, 몇 퍼센트를 샘플링 할 것인가?

➡ 0~1 사이의 값으로 지정하거나
n이라는 인자를 통해 절대 수치로 지정 가능

② Pandas Cheat Sheet로 복습하기

색인 기능 (Subset Observations(Rows))

✓ df.nlargest(n, 'value')

◉ 특정 컬럼의 값을 기준으로 상위 몇 개의 값을 가져옴

✓ df.nsmallest(n, 'value')

◉ 특정 컬럼의 값을 기준으로 하위 몇 개의 값을 가져옴

② Pandas Cheat Sheet로 복습하기

색인 기능 (Subset Observations(Rows))

✓ df.nlargest(n, 'value')

```
df
```

	population	GDP	alpha-2
Italy	59000000	1937894	IT
France	65000000	2583560	FR
Malta	434000	12011	MT
Maldives	434000	4520	MV
Brunei	434000	12128	BN
Iceland	337000	17036	IS
Nauru	11300	182	NR
Tuvalu	11300	38	TV
Anguilla	11300	311	AI

```
df.nlargest(3, 'population')
```

	population	GDP	alpha-2
France	65000000	2583560	FR
Italy	59000000	1937894	IT
Malta	434000	12011	MT

◉ population에서 상위 3개 리턴하는 함수

② Pandas Cheat Sheet로 복습하기

색인 기능 (Subset Variables (Columns))

- ✓ 기본적으로 columns을 먼저 색인함
- ✓ 여러 개를 색인하고 싶은 경우 : 리스트의 형태
  - ◉ df[['width','length','species']]
- ✓ 특정 columns을 지정할 경우
  - ◉ df['width'] or df.width
- ✓ 특정 columns과 특정 row를 동시에 색인하고 싶을 경우
  - ◉ df.loc[:,'x2':'x4']
  - ◉ df.iloc[:,[1,2,5]]
  - ◉ df.loc[df['a'] 〉 10, ['a','c']]

② Pandas Cheat Sheet로 복습하기

색인 기능 (Subset Variables (Columns))

- ✓ 기본적으로 columns을 먼저 색인함
- ✓ 여러 개를 색인하고 싶은 경우 : 리스트의 형태
  - ◉ df[['width','length','species']]
- ✓ 특정 columns을 지정할 경우
  - ◉ df['width'] or df.width
- ✓ 특정 columns과 특정 row를 동시에 색인하고 싶을 경우

이 형태는 첫 번째 인자가 로우 인덱스에 대한 정보,
두 번째 인자가 컬럼 인덱스에 대한 정보를 지정해 주시면 원하는 데이터만 색인해 올 수 있었습니다.

② Pandas Cheat Sheet로 복습하기

색인 기능 (Subset Variables (Columns))

✓ df.filter(regex='regex')

regex (Regular Expressions) Examples	
'\.'	Matches strings containing a period '.'
'Length$'	Matches strings ending with word 'Length'
'^Sepal'	Matches strings beginning with the word 'Sepal'
'^x[1-5]$'	Matches strings beginning with 'x' and ending with 1,2,3,4,5
'^(?!Species$).*'	Matches strings except the string 'Species'

② Pandas Cheat Sheet로 복습하기

색인 기능 (Subset Variables (Columns))

✓ df.filter(regex='regex')

```
df = pd.DataFrame(np.array(([1, 2, 3], [4, 5, 6])),
                  index=['mouse', 'rabbit'],
                  columns=['one', 'two', 'three'])
df
```

	one	two	three
mouse	1	2	3
rabbit	4	5	6

◉ 원하는 데이터를 필터(색인)하고자 함

② Pandas Cheat Sheet로 복습하기

색인 기능 (Subset Variables (Columns))

✓ df.filter(regex='regex')

	one	two	three
mouse	1	2	3
rabbit	4	5	6

```
df.filter(items=['one', 'three'])
```

	one	three
mouse	1	3
rabbit	4	6

굳이 filter를 사용할 필요가 없음

② Pandas Cheat Sheet로 복습하기

색인 기능 (Subset Variables (Columns))

✓ df.filter(regex='regex')

◉ 정규 표현식으로 컬럼명을 추출해 낼 수 있음

```
# e로 끝나는 컬럼들만 선택
df.filter(regex='e$', axis=1)
```

	one	three
mouse	1	3
rabbit	4	6

② Pandas Cheat Sheet로 복습하기

색인 기능 (Subset Variables (Columns))

✓ df.filter(regex='regex')

◉ 특정 문자열을 포함하는 컬럼을 선택할 때 사용

```
# bbi를 포함하는 행 선택
df.filter(like='bbi', axis=0)
```

	one	two	three
rabbit	4	5	6

axis=0 해당 문자열을 포함하고 있는 row 선택

axis=1 해당 문자열을 포함하고 있는 컬럼 선택

② Pandas Cheat Sheet로 복습하기

색인 기능 (Subset Variables (Columns))

✓ df.filter(regex='regex')

◉ 특정 문자열을 포함하는 컬럼을 선택할 때 사용

```
# bbi를 포함하는 행 선택
df.filter(like='bbi', axis=0)
```

	one	two	three
rabbit	4	5	6

axis=0 해당 문자열을 포함하고 있는 row 선택

서브스트링 매칭을 하거나 정규 표현식으로 특정 문자의 패턴을 통해서
로우나 컬럼을 필터링, 색인하고 싶은 경우에는 필터 함수를 사용할 수 있습니다.

② Pandas Cheat Sheet로 복습하기

색인 기능 (Subset Variables (Columns))

✓ df.filter(regex='regex')

regex (Regular Expressions) Examples	
'\.'	Matches strings containing a period '.'
'Length$'	Matches strings ending with word 'Length'
'^Sepal'	Matches strings beginning with the word 'Sepal'
'^x[1-5]$'	Matches strings beginning with 'x' and ending with 1,2,3,4,5
'^(?!Species$).*'	Matches strings except the string 'Species'

특정 문자열로 끝나거나 특정 문자열이 포함되어 있거나
혹은 특정 문자열로 시작하는 경우에 주로 많이 사용하게 됩니다.

② Pandas Cheat Sheet로 복습하기

데이터를 집계하는 함수 (Summarize Data)

✓ df['w'].value_counts( )

◉ 특정 컬럼의 값들의 카운트하는 함수

예

100명의 학생들이 1, 2, 3반에 할당되어 있는 경우

각 반이 몇 명이 있는지 통계

② Pandas Cheat Sheet로 복습하기

데이터를 집계하는 함수 (Summarize Data)

✓ len(df)

◉ 데이터프레임에 있는 row의 개수를 리턴

✓ df['w'].nunique( )

◉ 유일한 값의 개수

예

100명의 학생들이 1, 2, 3반에 할당되어 있는 경우

총 몇 개의 반이 있는지, 즉 3을 리턴

② Pandas Cheat Sheet로 복습하기

데이터를 집계하는 함수 (Summarize Data)

✓ df.describe( )

◉ 데이터프레임의 기본 통계 정보를 보여주는 함수

✓ 그 외

◉ sum( ) ◉ count( ) ◉ median( )

◉ min( ) ◉ max( ) ◉ var( ) ◉ std( )

② Pandas Cheat Sheet로 복습하기

7 새로운 컬럼을 생성하는 방법 (Make New Columns)

1. 기존에 있는 컬럼값들을
   산술 연산을 해서 새로운 컬럼 생성

2. 내가 만든 규칙을 적용한 새로운 컬럼 생성

3. 내가 만든 규칙을 적용하기 위해서는
   내가 만든 규칙을 어떤 함수로 정의하고
   그 함수를 어떤 컬럼에 적용해서 적용된 값을
   새로운 컬럼으로 추가

② Pandas Cheat Sheet로 복습하기

7 새로운 컬럼을 생성하는 방법 (Make New Columns)

✓ pd.qcut(df.col, n, labels=False)

```
#1~500 사이의 값 100개 생성
data = pd.Series(random.sample(range(1,500), 100))
data
```

```
0      154
1      269
2      308
3      333
4      247
      ...
95      87
96     372
97     449
98     254
99     381
Length: 100, dtype: int64
```

② Pandas Cheat Sheet로 복습하기

7 새로운 컬럼을 생성하는 방법 (Make New Columns)

✓ pd.qcut(df.col, n, labels=False)

- ◉ data ➡ qcut하고자 하는 데이터
- ◉ 4 ➡ 나눌 데이터의 개수
- ◉ 1, 2, 3, 4
  ➡ 나눠진 각각의 그룹들의 이름 지정

```
pd.qcut(data, 4, labels = [1,2,3,4])
```

```
0      2
1      3
2      3
3      3
4      3
      ..
95     1
96     4
97     4
98     3
99     4
Length: 100, dtype: category
```

qcut은 데이터를 동일한 크기의 4개의 그룹으로 분할하고
각 그룹의 이름을 1, 2, 3, 4로 부여해서 결과를 리턴해 주게 됩니다.

② Pandas Cheat Sheet로 복습하기

7 새로운 컬럼을 생성하는 방법 (Make New Columns)

✓ pd.qcut(df.col, n, labels=False)

◉ 4개의 그룹으로 분할하는 기준은
각각의 그룹의 개수가 동일해야 함

```
pd.qcut(data, 4, labels = [1,2,3,4]).value_counts()
```

```
4    25
3    25
2    25
1    25
```

qcut()은 내가 정해 놓은 그룹의 개수만큼 데이터를 그룹핑하며,
각 그룹의 크기는 동일하게 만드는 함수입니다.

② Pandas Cheat Sheet로 복습하기

Group Data

✓ df.groupby(by="col")

df.groupby(by='col') ⊕ agg(function)

↓

df.groupby('col').agg('sum')

df.groupby('col'). sum()

② Pandas Cheat Sheet로 복습하기

Group Data

✓ df.pivot_table(index = 'col1',
columns = 'col2',
aggfunc = 'sum',
values = 'col3')

◉ index, columns : 그룹핑하고자 하는 컬럼명

◉ aggfunc, values : 집계하고자 하는 함수와 함수를 적용하고자 하는 컬럼 지정

② Pandas Cheat Sheet로 복습하기

데이터셋 합치기 (Combine Data Sets)

✓ pd.merge( )

② Pandas Cheat Sheet로 복습하기

데이터셋 합치기 (Combine Data Sets)

✓ pd.merge( )

- sql의 join 함수의 연산과 동일
- 2개의 데이터프레임을 합침

② Pandas Cheat Sheet로 복습하기

데이터셋 합치기 (Combine Data Sets)

✓ pd.merge( )

예

```
pd.merge(adf, bdf, how='left', on='x1')
```

adf

x1	x2
A	1
B	2
C	3

+

bdf

x1	x3
A	T
B	F
D	T

=

x1	x2	x3
A	1	T
B	2	F
D	3	NaN

◉ on 인자 : 명시적으로 지정하지 않으면 공통된 columns으로 자동 설정

② Pandas Cheat Sheet로 복습하기

데이터셋 합치기 (Combine Data Sets)

✓ pd.merge( )

◉ how 인자 : left, right, outer, inner

➡ pd.merge(adf, bdf, how='inner', on='x1')

x1	x2	x3
A	1	T
B	2	F

② Pandas Cheat Sheet로 복습하기

데이터셋 합치기 (Combine Data Sets)

✓ pd.merge( )

◉ how 인자 : left, right, outer, inner

➡ pd.merge(adf, bdf, how='left', on='x1')

➡ pd.merge(adf, bdf, how='right', on='x1')

➡ pd.merge(adf, bdf, how='outer', on='x1')

x1	x2	x3
A	1	T
B	2	F

inner 결과

left에 있던 데이터프레임 중
실제 결과에 포함되어
있지 않은 값

② Pandas Cheat Sheet로 복습하기

데이터셋 합치기 (Combine Data Sets)

✓ pd.merge( )

◉ how 인자 : left, right, outer, inner

→ outer 예

adf

x1	x2
A	1
B	2
C	3

⊕

bdf

x1	x3
A	T
B	F
D	T

=

x1	x2	x3
A	1	T
B	2	F
C	3	NaN
D	NaN	T

Pandas Cheat Sheet 한장으로 Pandas 복습하기

학습완료

1/ Pandas Cheat Sheet 소개

2/ Pandas Cheat Sheet로 복습하기

수고하셨습니다!