3. pandas 시작하기

1강. pandas 라이브러리

학습내용

- pandas의 개념
- 데이터 프레임 생성
- 데이터 프레임의 기초

학습목표

- pandas의 개념에 대해 설명할 수 있다.
- pandas로 데이터 프레임을 생성하고 불러올 수 있다.
- pandas에서 사용되는 자료형을 구별할 수 있다.

1. pandas의 개념

1) 기본 문법

가. 정의

- pandas는 데이터 분석용 라이브러리로 데이터를 다루는 패키지 중 하나
- 데이터 분석을 위한 효율적인 데이터 구조를 제공하며, 1차원 배열 형태의 데이터 구조인 Series와 2차원 배열 형태의 데이터 구조인 Data Frame

나. 특징

- pandas는 파이썬에서 가장 널리 사용되는 데이터 분석 라이브러리로 데이터 프레임(DataFrame) 이라는 자료구조
- 데이터 프레임은 엑셀의 스프레드시트와 유사한 형태이며 파이썬으로 데이터를 쉽게 처리할 수 있음
- 데이터를 분석 및 조작을 위한 라이브러리
- 수치형 테이블과 시계열 데이터를 조작하고 운영하기 위한 데이터를 제공
- 시계열 데이터와 비시계열 데이터를 함께 다룰 수 있는 통합 자료 구조
- 누락된 데이터를 유연하게 처리할 수 있는 기능
- SQL 같은 일반 데이터베이스처럼 데이터를 합치고 관계연산을 수행하는 기능

다. 라이브러리

- pandas 라이브러리는 내장 라이브러리가 아니기 때문에 원래는 별도로 설치해야 하지만 아나콘다 배포판을 사용하는 경우 내장되어 있어 따로 설치할 필요 없으며 만약 파이썬 IDLE를 사용하는 경우 명령 프롬프트에서 'pip install' 옆에 설치하고 싶은 라이브러리 이름을 입력하여 필요한 라이브러리 설치할 수 있음.

2) 데이터 프레임 생성

1) 기본

가. 라이브러리 추가

- 데이터 분석 라이브러리를 import하는 코드로 pandas 라이브러리는 보통 pd로 축약

  import pandas as pd

나. 예제

```
# 판다스의 데이터 프레임을 생성
name = ['Bob', 'Jessica', 'Mary', 'John', 'Mel']
births = [968, 155, 77, 578, 973]
custom = [1, 5, 25, 13, 23232]

DataSet = list(zip(names, births))
df = pd.DataFrame(data = DataSet,
columns=['Names', 'Births'])
# 데이터 프레임의 상단 부분을 출력함
df.head()
```

2) pandas 자료형

가. Series

- 시트의 열 1개를 의미
- 1차원 배열 형태의 데이터 구조를 사용
- 가로 방향으로 크기 변경
- 색인 추가 가능

- 예제

```
# index와 value를 가짐
# 색인(index)을 지정할 수 있음
import pandas as pd
series_ex = pd.Series ( [ '월' , '화', '수', '목', '금' ] ,
                        index=[0, 1, 2, 3, 4])
print(series_ex)
```

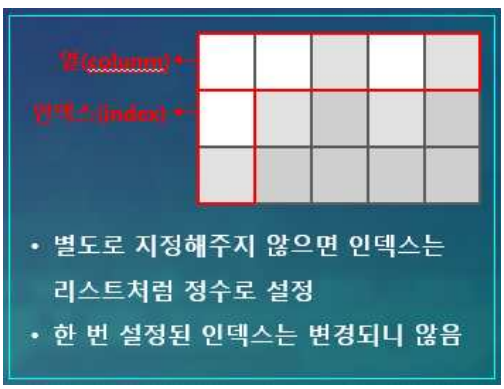
결과

```
0    월
1    화
2    수
3    목
4    금
dtype: object
```

pandas는 문자열 자료형을 object라는 이름으로 인식하고 파이썬은 string이라 인식

나. DataFrame

- 엑셀에서 볼 수 있는 시트(Sheet)와 동일한 개념
- 2차원 배열 형태의 데이터 구조로 행과 열이 있음
- 행을 구분해주는 인덱스(index)와 열을 구분해주는 컬럼(column) 있음



- 예제

```
# 판다스의 데이터 프레임을 생성
import pandas as pd
name = ['Bob', 'Jessica', 'Mary', 'John', 'Mel']
births = [968, 155, 77, 578, 973]
custom = [1, 5, 25, 13, 23232]

DataSet = list(zip(names, births))
df = pd.DataFrame(data = DataSet,
columns=['Names', 'Births'])
# 데이터 프레임의 상단 부분을 출력함
df.head()
```

결과

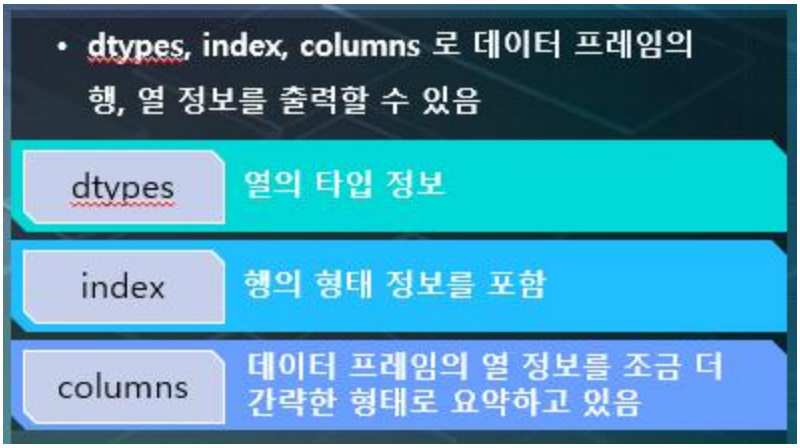
	Names	Births
0	Bob	968
1	Jessica	155
2	Mary	77
3	John	578
4	Mel	973

- 표와 같은 스트레드시트 형식(ex:엑셀)의 자료구조
- 칼럼(column)으로 구성되어 있으며, 숫자, 문자열, 논리형(True/False)으로 저장
- 경우에 따라 DataFrame은 df로 축약하여 사용

3) 데이터 프레임

1) 문법

가. 기본 구조

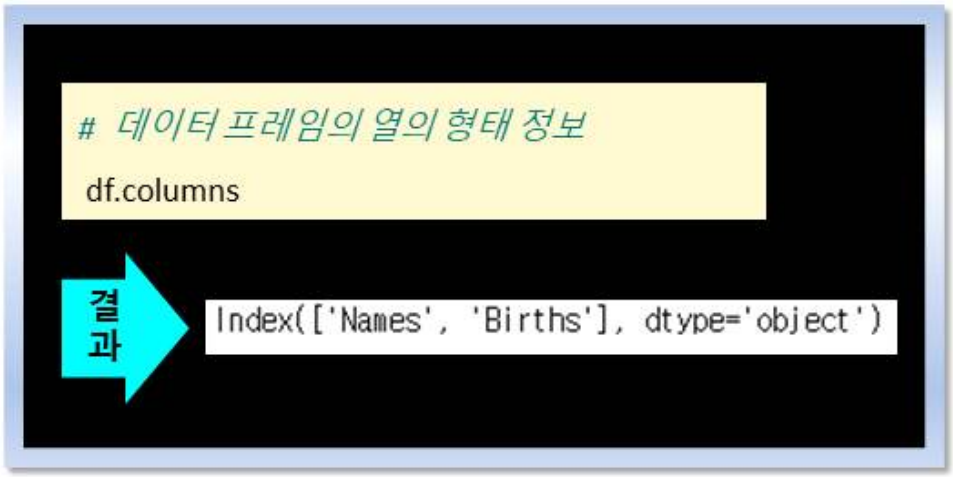


나. dtypes – 열의 타입 정보

다. pandas자료형과 파이썬 자료형 비교

라. index – 행의 형태 정보

마. columns - 데이터프레임의 열 정보



마. 데이터 프레임의 열 선택하기

사. 데이터 프레임의 인덱스 선택하기


0 ~ 3 번째 인덱스를 선택
df[0 : 3]
결과
Names Births
0 Bob 968
1 Jessica 155
2 Mary 77

평가하기

1. pandas의 주요특징에 해당하지 않는 것은?

① 자동적/명시적으로 축의 이름에 따라 데이터를 정렬 가능

② 잘못 정렬된 데이터에 의한 오류를 방지하지 못함

③ 시계열 데이터와 비시계열 데이터를 함께 다룰 수 있는 통합 자료 구조

④ 누락된 데이터를 유연하게 처리할 수 있는 기능

⑤ SQL 같은 일반 데이터베이스처럼 데이터를 합치고 관계연산을 수행하는 기능

- 정답 : ②번

해설 : pandas는 잘못 정렬된 데이터에 의한 오류를 방지하는 기능이 포함됩니다.

2. pandas 사용시 필요한 라이브러리 추가 구문은?

① Import pd

② Import pandas as pd

③ import pandas

④ import PANDAS as pd

⑤ Import 판다스

- 정답 : ③

해설 : 라이브러리를 추가할 경우 import pandas 라고 사용하고 pd로 줄여 사용할 경우 as pd를 추가합니다.(대소문자 구분!)

학습정리

1. pandas

- pandas는 파이썬에서 가장 널리 사용되는 데이터 분석 라이브러리로 데이터 프레임(DataFrame) 이라는 자료구조

2. 데이터 프레임 생성

- 데이터 프레임(DataFrame)은 엑셀에서 볼 수 있는 시트(Sheet)와 동일한 개념
- 시리즈(Series)는 시트의 열 1개를 의미

3. 데이터 프레임의 기초

- dtypes - 열의 타입 정보
- index - 행의 형태 정보를 포함
- columns - 데이터프레임의 열 정보