파이썬 머신러닝 판다스 데이터분석

Lecture (2)

Dr. Heesuk Kim

목차

Part 1. 판다스 입문

1. 데이터과학자가 판다스를 배우는 이유

빅데이터의 시대. 데이터 과학이라는 새로운 영역의 출현.

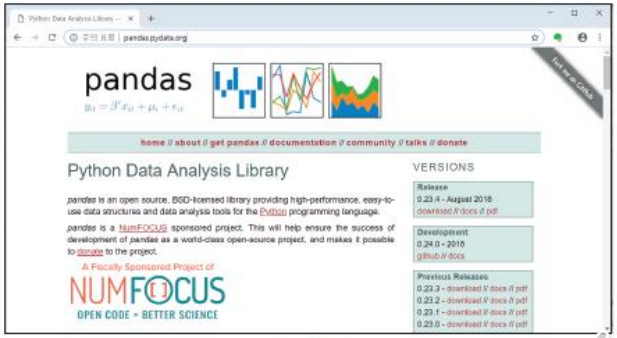
- 클라우드 컴퓨팅의 확산. 빅데이터 저장, 분석에 필요한 컴퓨팅 자원이 매우 저렴해짐.
- 컴퓨팅 파워의 대중화는 최적의 학습환경과 연구 인프라를 제공.

데이터과학은 데이터를 연구하는 분야이고, 데이터 자체가 가장 중요한 자원

- 데이터 분석 업무의 80~90%는 데이터를 수집하고 정리하는 일이 차지.
- 나머지 10~20%는 알고리즘을 선택하고, 모델링 결과를 분석하여 데이터로부터 유용한 정보(information)를 뽑아내는 분석 프로세스의 몫.
- 데이터과학자가 하는 가장 중요한 일이 데이터를 수집하고 분석이 가능한 형태로 정리하는 것.

판다스는 데이터를 수집하고 정리하는데 최적화된 도구.

- 가장 배우기 쉬운 프로그래밍 언어, 파이썬(Python) 기반.
- 오픈소스(open source)로 무료로 이용 가능.



[그림 1-1] 판다스 공식 홈페이지(http://pandas.pydata.org/)

2. 판다스 자료구조

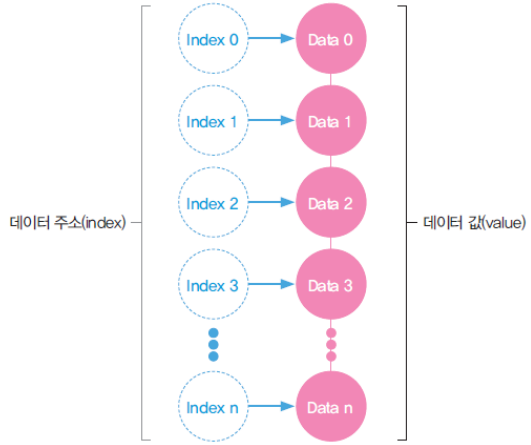
- **목적**

분석을 위해 다양한 소스(source)로부터 수집하는 데이터는 형태나 속성이 매우 다양하다. 특히, 서로 다른 형식을 갖는 여러 종류의 데이터를 컴퓨터가 이해할 수 있도록 동일한 형식을 갖는 구조로 통합할 필요가 있다. 판다스의 일차적인 목적은 **형식적으로 서로 다른 여러 가지 유형의 데이터를 공통의 포맷으로 정리**하는 것이다.

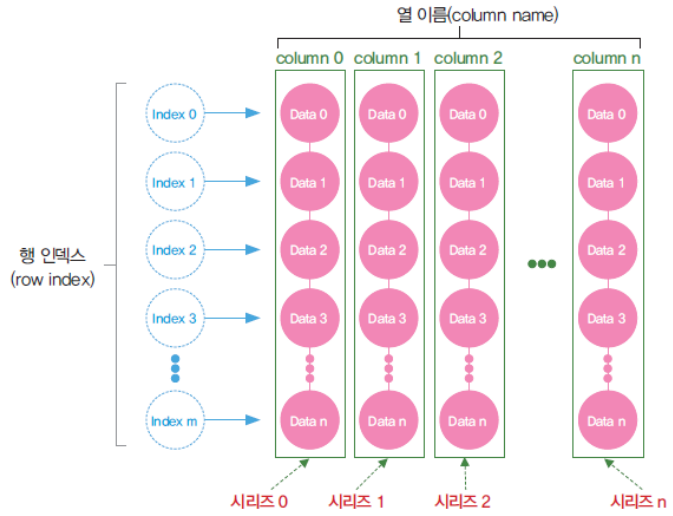
- **종류**

판다스는 **시리즈(Series)와 데이터프레임(DataFrame)이라는 구조화된 데이터 형식**을 제공한다. 서로 다른 종류의 데이터를 한곳에 담는 그릇(컨테이너)이 된다. 다만, 시리즈는 1차원 배열이고, 데이터프레임이 2차원 배열이라는 점에서 차이가 있다. 특히, 행과 열로 이루어진 2차원 구조의 데이터프레임은 데이터 분석 실무에서 자주 사용된다.

2. 판다스 자료구조

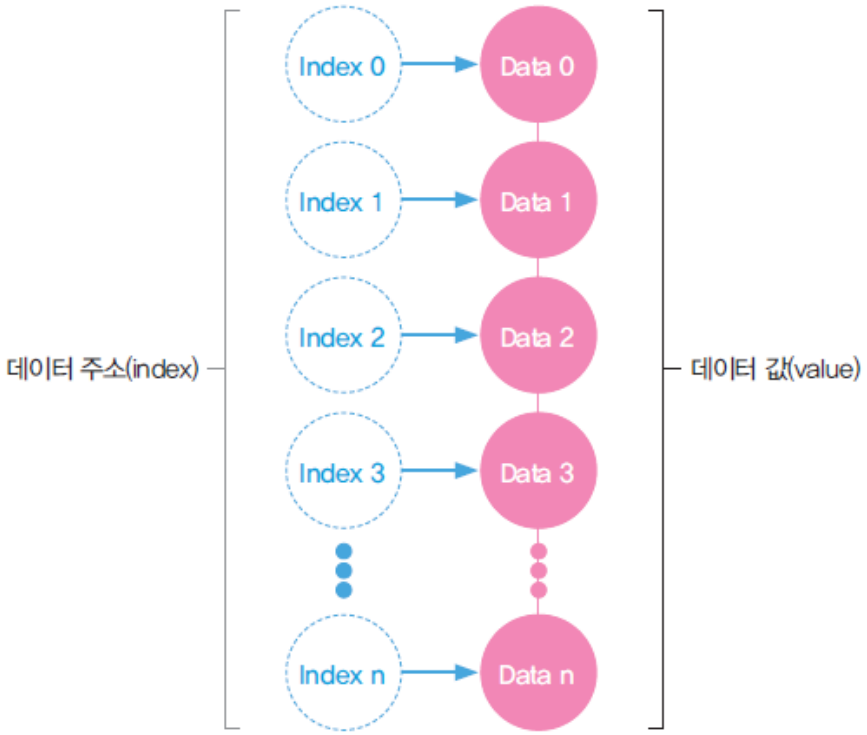


[그림 1-2] 시리즈 구조



[그림 1-6] 데이터프레임 구조

2. 판다스 자료구조



[그림 1-2] 시리즈 구조

2-1. 시리즈

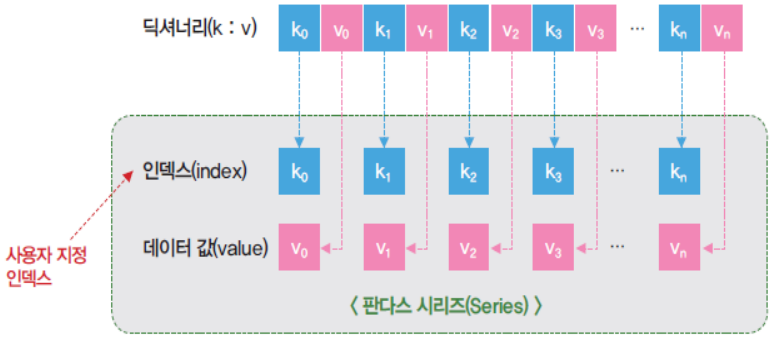
- **시리즈 만들기**

1) 딕셔너리와 시리즈의 구조가 비슷하기 때문에, **딕셔너리를 시리즈로 변환**하는 방법을 많이 사용.

2) 판다스 내장 함수인 Series()를 이용하고, 딕셔너리를 함수의 매개변수(인자)로 전달.

```
딕셔너리 → 시리즈 변환: pandas.Series( 딕셔너리 )
```

3) 딕셔너리의 키(k)는 시리즈의 인덱스에 대응하고, 딕셔너리의 각 키에 매칭되는 값(v)이 시리즈의 데이터 값(원소)로 변환.



[그림 1-3] 딕셔너리 → 시리즈 변환

2-1. 시리즈

예제 1-1

딕셔너리를 시리즈로 변환해 본다. {'a': 1, 'b': 2, 'c': 3}와 같이 'k:v' 구조를 갖는 딕셔너리를 정의하여 변수 dict_data에 저장한다. 변수 dict_data에 저장되어 있는 딕셔너리를 Series() 함수의 인자로 전달하면, 시리즈로 변환한다. Series() 함수가 반환한 시리즈 객체를 변수 sr에 저장한다.

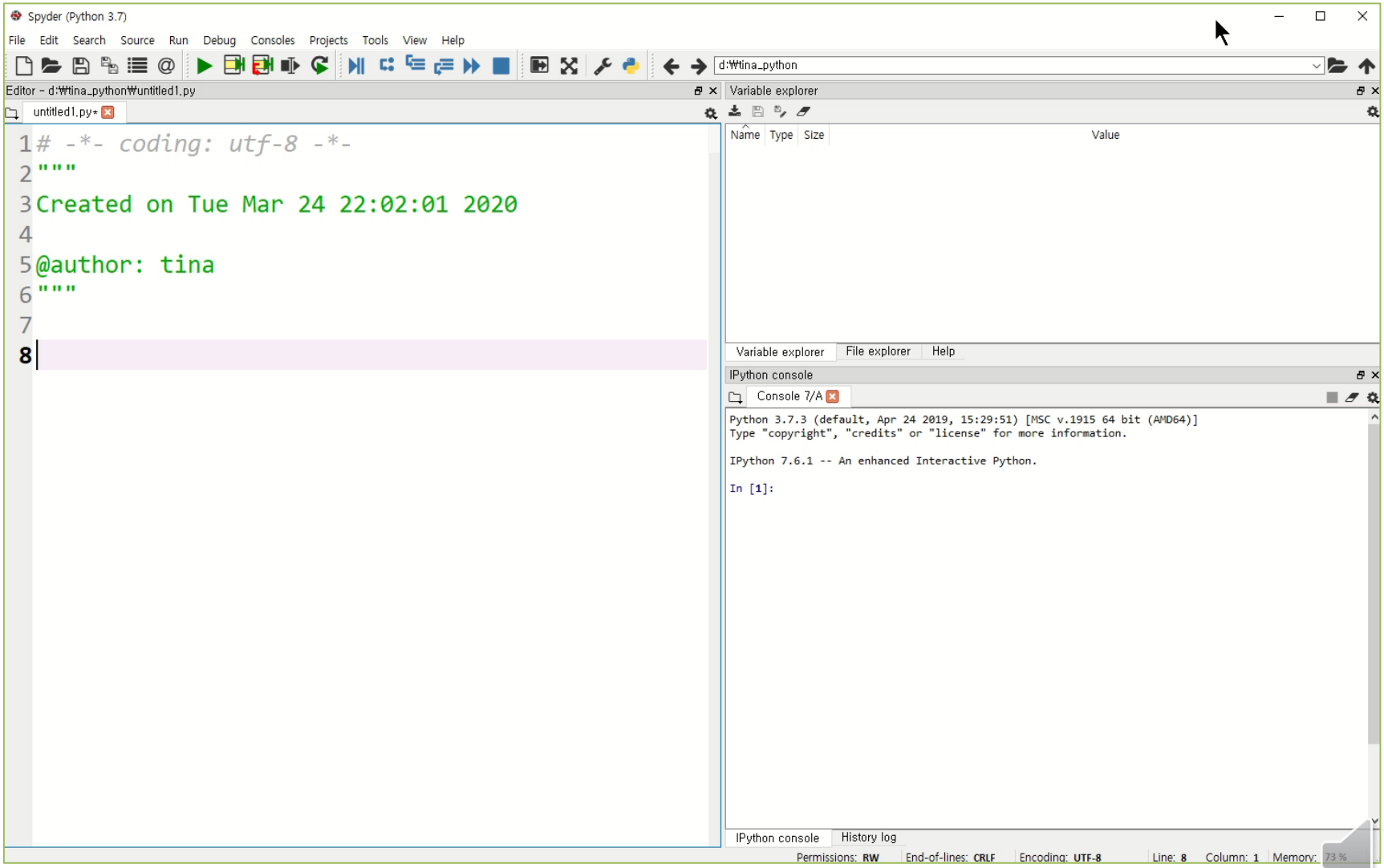
〈예제 1-1〉 딕셔너리 → 시리즈 변환 (File: example/part1/1.1_dict_to_series.py)

```
1  # -*- coding: utf-8 -*-
2
3  # pandas 불러오기
4  import pandas as pd
5
6  # key:value 쌍으로 딕셔너리를 만들고, 변수 dict_data에 저장
7  dict_data = {'a': 1, 'b': 2, 'c': 3}
8
9  # 판다스 Series() 함수로 dictionary를 Series로 변환. 변수 sr에 저장
10 sr = pd.Series(dict_data)
11
12 # sr의 자료형 출력
13 print(type(sr))
14 print('\n')
15 # 변수 sr에 저장되어 있는 시리즈 객체를 출력
16 print(sr)
```

〈실행 결과〉 코드 전부 실행

```
<class 'pandas.core.series.Series'>


a    1
b    2
c    3
dtype: int64
```

2-1. 시리즈 예제 1-1


Spyder (Python 3.7)
File Edit Search Source Run Debug Consoles Projects Tools View Help
d:₩tina_python
Editor - d:₩tina_python₩untitled1.py
untitled1.py
1 # -*- coding: utf-8 -*-
2 """
3 Created on Tue Mar 24 22:02:01 2020
4
5 @author: tina
6 """
7
8
Variable explorer
Name Type Size Value
Variable explorer File explorer Help
IPython console
Console 7/A
Python 3.7.3 (default, Apr 24 2019, 15:29:51) [MSC v.1915 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.
IPython 7.6.1 -- An enhanced Interactive Python.
In [1]:
IPython console History log
Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 8 Column: 1 Memory: 73 %

2-1. 시리즈

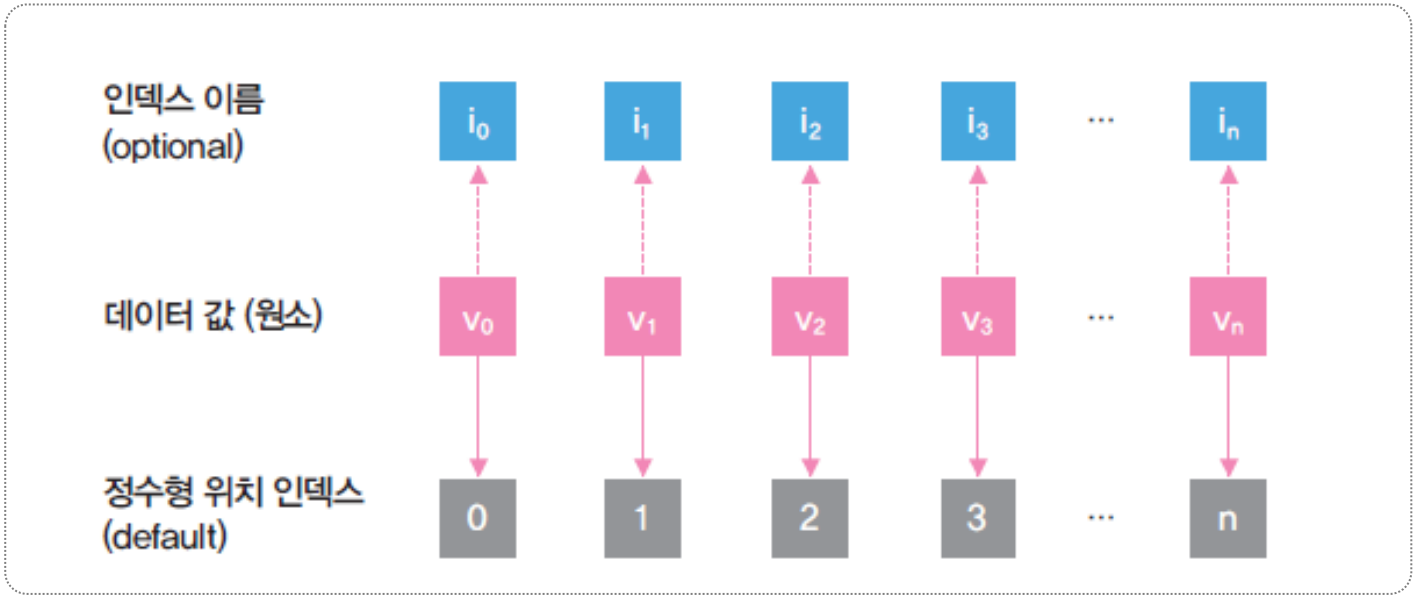
- **인덱스 구조**

1) 인덱스는 자기와 짝을 이루는 **원소의 순서와 주소를 저장**.

2) 인덱스를 잘 활용하면 데이터 값의 탐색, 정렬, 선택, 결합 등 데이터 조작을 쉽게 할 수 있다.

3) **인덱스의 종류 (2가지)**

①정수형 위치 인덱스(integer position)
②인덱스 이름(index name) 또는 인덱스 라벨(index label)



[그림 1-4] 시리즈 인덱스의 유형

2-1. 시리즈

예제 1-2

① 시리즈 만들기

파이썬 리스트를 시리즈로 변환해 본다. 단, 딕셔너리의 키처럼 인덱스로 변환될 값이 없다. 따라서, 인덱스를 별도로 정의하지 않으면, 디폴트로 정수형 위치 인덱스(0, 1, 2, …)가 자동 지정된다. 다음 예제에서는 0 ~ 4 범위의 정수값이 인덱스로 지정된다.

〈예제 1-2〉 시리즈 인덱스 (File: example/part1/1.2_series_index.py)

```
1  # -*- coding: utf-8 -*-
2
3  import pandas as pd
4
5  # 리스트를 시리즈로 변환하여 변수 sr에 저장
6  list_data = ['2019-01-02', 3.14, 'ABC', 100, True]
7  sr = pd.Series(list_data)
8  print(sr)
```

〈실행 결과〉 코드 1~8라인을 부분 실행

```
0    2019-01-02
1          3.14
2           ABC
3           100
4          True
dtype: object
```

2-1. 시리즈

예제 1-2

② 인덱스 vs. 데이터 값 배열 확인하기

시리즈의 index 속성과 values 속성을 이용하면, 인덱스 배열과 데이터 값의 배열을 불러올 수 있다. .

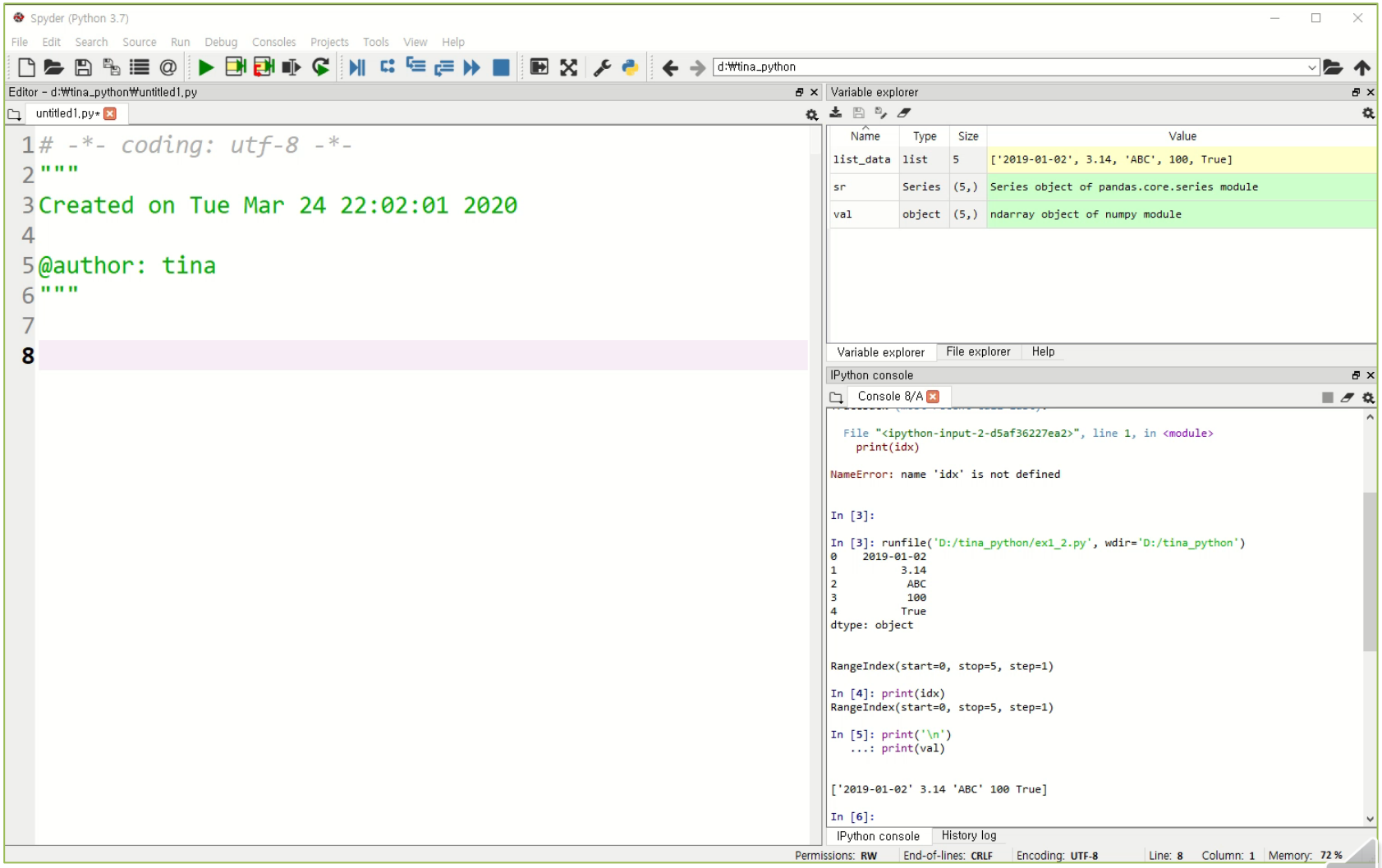
〈예제 1-2〉 시리즈 인덱스 (File: example/part1/1.2_series_index.py(이어서 계속))

```
~  ~~~ 생략 ~~~

11  # 인덱스 배열은 변수 idx에 저장. 데이터 값 배열은 변수 val에 저장
12  idx = sr.index
13  val = sr.values
14  print(idx)
15  print('\n')
16  print(val)
```

〈실행 결과〉 코드 11~16라인을 부분 실행

```
RangeIndex(start=0, stop=5, step=1)


['2019-01-02' 3.14 'ABC' 100 True]
```

2-1. 시리즈 예제 1-2


Spyder (Python 3.7)
File Edit Search Source Run Debug Consoles Projects Tools View Help
d:₩tina_python
Editor - d:₩tina_python₩untitled1.py
untitled1.py*
1 # -*- coding: utf-8 -*-
2 """
3 Created on Tue Mar 24 22:02:01 2020
4
5 @author: tina
6 """
7
8
Variable explorer
Name Type Size Value
list_data list 5 ['2019-01-02', 3.14, 'ABC', 100, True]
sr Series (5,) Series object of pandas.core.series module
val object (5,) ndarray object of numpy module
Variable explorer File explorer Help
IPython console
Console 8/A
File "<ipython-input-2-d5af36227ea2>", line 1, in <module>
print(idx)
NameError: name 'idx' is not defined
In [3]:
In [3]: runfile('D:/tina_python/ex1_2.py', wdir='D:/tina_python')
0 2019-01-02
1 3.14
2 ABC
3 100
4 True
dtype: object
RangeIndex(start=0, stop=5, step=1)
In [4]: print(idx)
RangeIndex(start=0, stop=5, step=1)
In [5]: print('\n')
...: print(val)
['2019-01-02' 3.14 'ABC' 100 True]
In [6]:
IPython console History log
Permissions: RW End-of-lines: CRLF Encoding: UTF-8 Line: 8 Column: 1 Memory: 72 %

Any Question?

idishskim@naver.com

Thank you.