본 페이지는 편집 시 유의사항을 담고 있습니다.

페이지 수 오류를 방지하기 위해 실제 편집 작업 시에는 지워 주시기 바랍니다.

이 파일은 B5로 제작을 원하는 저자분들께서 사용하시면 됩니다.

편집 과정중 꽉차게 들어가는 이미지, 배경색상이 있을 경우 하얀색 종이가 보이지 않도록 이미지, 색상 크기 반영하셔야 하는 점 알려드립니다.

이파일은 실제 제작시 필요한 사방여백 3mm가 포함된 크기입니다.(188x263mm)

상하좌우 3mm는 실제 제작시 재단되어 반영되지 않으니 참고해주세요!

**어린 왕자**(제목을 적어주세요.)

**어린 왕자(제목을 적어주세요)**

**발 행**｜2023년 00월 00일

**저 자**｜생텍쥐 페리(저자명, 필명을 적어주세요)

**펴낸이**｜한건희

**펴낸곳**｜주식회사 부크크

**출판사등록**｜2014.07.15(제2014-16호)

**주 소**｜서울특별시 금천구 가산디지털1로 119 SK트윈타워 A동 305호

**전 화**｜1670-8316

**이메일**｜info@bookk.co.kr

**ISBN**｜979-11-410-0000-0

**www.bookk.co.kr**

**ⓒ 셍텍쥐 페리 2023**

본 책은 저작자의 지적 재산으로서 무단 전재와 복제를 금합니다.

어

린

왕

자

Evan, Sarah 지음

**CONTENT**

내용

[Chapter 1. 집필 동기 9](#_Toc130723838)

[1. 이 책을 쓰게 된 동기 9](#_Toc130723839)

[A. 강사 입장에서 9](#_Toc130723840)

[B. 수강생 입장에서 9](#_Toc130723841)

[2. 무엇을 배울 수 있는가 9](#_Toc130723842)

[Chapter 2. 개발환경설정 10](#_Toc130723843)

[1. Github 회원가입 10](#_Toc130723844)

[2. Streamlit 회원가입 10](#_Toc130723845)

[3. Git 설치 10](#_Toc130723846)

[4. Python 설치 10](#_Toc130723847)

[5. Visual Studio Code 설치 10](#_Toc130723848)

[6. 가상환경 Virtualenv 설치 10](#_Toc130723849)

[7. requirements.txt 파일 작성 10](#_Toc130723850)

[8. Streamlit, Hello World 10](#_Toc130723851)

[Chapter 3. Python 기본문법 11](#_Toc130723852)

[1. Python Basic Tutorial 11](#_Toc130723853)

[A. Python 프로그래밍 시작 11](#_Toc130723854)

[B. 주석처리 11](#_Toc130723855)

[C. Variables 12](#_Toc130723856)

[D. 사칙연산 15](#_Toc130723857)

[E. 비교 연산자Comparison Operators 16](#_Toc130723858)

[F. 논리 연산자Logical Operators 17](#_Toc130723859)

[G. String 19](#_Toc130723860)

[2. Pandas 25](#_Toc130723861)

[3. Matplotlib & Seaborn 25](#_Toc130723862)

[4. Plotly 25](#_Toc130723863)

[5. Scikit-Learn 25](#_Toc130723864)

[Chapter 4. Streamlit 26](#_Toc130723865)

[1. Slider bar 26](#_Toc130723866)

[2. Select bar 26](#_Toc130723867)

[3. Tab 26](#_Toc130723868)

[4. Checkbox 26](#_Toc130723869)

[Chapter 5. 공공데이터 수집 27](#_Toc130723870)

[1. 서울열린데이터광장 27](#_Toc130723871)

[2. 공공데이터포털 27](#_Toc130723872)

[Chapter 6. 부동산 실거래가 대시보드 28](#_Toc130723873)

[1. 데이터 수집 28](#_Toc130723874)

[2. Home 화면 꾸미기 28](#_Toc130723875)

[3. 탐색적 자료분석 화면 꾸미기 28](#_Toc130723876)

[4. 머신러닝 모형 만들기 28](#_Toc130723877)

[5. Prediction 화면 꾸미기 28](#_Toc130723878)

[6. OpenAI 챗봇 구현하기 28](#_Toc130723879)

[Chapter 7. 배포 29](#_Toc130723880)

[1. secrets.toml 29](#_Toc130723881)

[2. 배포 29](#_Toc130723882)

[3. 마무리 29](#_Toc130723883)

[인용 자료 30](#_Toc130723884)

작가의 말 242

레옹 베르트에게 이 책을 바칩니다.

이 책을 어른에게 바친 데 대해 어린이들에게 용서를 바랍니다.

나에게는 그럴 만한 중요한 이유가 있습니다.

그것은 무엇보다도 그 사람은 이 세상에서 나와 가장 친한

친구이기 때문라는 점입니다. 그리고 그는 무엇이든지

알아들을 수 있으며 어린이들을 위한 책까지도 다 이해한다는 점입니다.

세 번째 이유는 그가 프랑스에 살고 있는데

그 곳에서 추위와 굶주림에 떨고 있다는 사실입니다.

그는 위로받아야 할 처지에 있는 것입니다.

그래도 이 모든 이유들이 부족하다면 예전 어린 시절의

그에게 이 책을 바치겠습니다. 어른들도 모두 한때는 어린이였으니까요.

(물론 그것을 기억하는 어른은 별로 없지만.)

그래서 바치는 글을 이렇게 고쳐 씁니다.

'어린 시절의 레옹 베르트에게 이 책을 바칩니다

# Chapter 1. 집필 동기

## 이 책을 쓰게 된 동기

### 강사 입장에서

### 수강생 입장에서

## 무엇을 배울 수 있는가

# Chapter 2. 개발환경설정

## Github 회원가입

### Streamlit 회원가입

### Git 설치

### Python 설치

### Visual Studio Code 설치

### 가상환경 Virtualenv 설치

### requirements.txt 파일 작성

### Streamlit, Hello World

# Chapter 3. Python 기본문법

## Python Basic Tutorial

### Python 프로그래밍 시작

프로그래밍의 첫번째 시작은 “Hello World!”를 출력하는 것에서부터 시작한다. Python에서는 다음과 같이 print()함수를 사용한다.

|  |
| --- |
| print("Hello, Streamlit!")  [결과]  Hello, Streamlit! |

### 주석처리

주석처리는 크게 1줄 주석처리와 다중 주석처리가 존재한다. 다음 코드를 통해 확인해 본다. 아래 코드를 실행하면 주석 처리된 곳은 생략하고, print( ) 함수만 실행되는 것을 확인할 수 있다.

|  |
| --- |
| # 한 줄 주석 처리  """  abc  다중 주석 처리  1234556  """  print("Hello, Streamlit!")  [결과]  Hello, Streamlit! |

### Variables

프로그래밍에서 변수Variables를 저장하는 방법은 크게 4가지(Camel case, Pascal case, Kebab case, Snake case) 방식으로 구분한다. 원 단어가 임의의 학생을 저장하는 변수 my student가 있다고 가정한다. 각자 원하는 방식을 사용하도록 한다. 다음 예시를 통해서 관련 코드를 익히도록 한다.

|  |
| --- |
| myStudent = “sarah” # Camel Case  MyStudent = “sarah” # Pascal Case  my-student = “sarah” # Kebab Case  my\_student = “sarah” # Snake Case |

Python에서 Data Structures를 말할 때, Primitive 자료형과 Non-Primitive 자료형으로 구분한다 (Jaiswal, 2017).[[1]](#footnote-1)

Table 1. Python Data Structures

|  |  |
| --- | --- |
| Primitive | Non-Primitive |
| Integer  Float  String  Boolean | Array  List  Tuple  Dictionary  Set  File |

여기에서 List는 크게 아래와 같이 Linear와 Non-Linear와 구분할 수 있다.

Table 2. List Data Types[[2]](#footnote-2)

|  |  |
| --- | --- |
| Linear | Non-Linear |
| Stacks  Queues | Graphs  Trees |

#### Integer

Integer는 정수형을 말하며, 4, 5, -1과 같은 숫자를 의미한다. 각자의 나이를 입력하고 출력하는 코드를 작성한다.

|  |
| --- |
| num\_int = 30  print(num\_int)  print(type(num\_int))  [결과]  30  <class 'int'> |

#### Float

Float는 실수형을 말하며, 일반적으로 1.11또는 3.11과 같은 소수를 표현한다. 이번에는 본인의 키를 입력하고 출력하는 코드를 작성한다.

|  |
| --- |
| my\_height = 181.8  print(my\_height)  print(type(my\_height))  [결과]  181.8  <class 'float'> |

#### String

String은 문자열을 말하며, 알파벳, 단어 또는 기타 문자의 모음이다. Python에서는 한쌍의 작은 따옴표(‘’) 또는 큰 따옴표(“”) 안에 일련의 문자를 포함하여 문자열을 생성할 수 있다. 우선 다음 코드를 확인해본다. 임의의 글자를 입력하고 출력하는 코드를 작성한다. 문자열을 다루는 방법에 대해서는 추후에 한번 더 다룰 예정이다.

|  |
| --- |
| my\_name = "evan"  print(my\_name)  print(type(my\_name))  [결과]  evan  <class 'str'> |

#### Boolean

일반적으로 True(참) 또는 False(거짓)을 표현하는 데이터 타입이며, True는 1로, False는 0으로 변환되기도 한다. 해당 데이터 타입은 특히 조건식 및 비교식에 매우 유용하게 활용할 수 있다. 임의의 두개의 숫자를 비교하는 코드를 작성하고, 결괏값을 확인해본다.

|  |
| --- |
| x = 5  y = 3  z = x == y  print(z)  print(type(z))  [결과]  False  <class 'bool'> |

### 사칙연산

일반적으로 형변환이라고 부른다. 임의의 두개를 변수를 만들고 사칙연산을 만들어 보도록 한다. 이 때 주의해서 봐야 할 것은 나눗셈이다. 두 개의 정수를 나눌 때, 정수가 실수형으로 변한다는 것에 주의해야 한다.

|  |
| --- |
| x = 6  y = 3  print(x + y)  print(x - y)  print(x \* y)  print(x / y)  [결과]  9  3  18  2.0 |

이번에는 사칙연산을 통해 나머지, 몫, 제곱을 구하는 연산자를 확인한다.

|  |
| --- |
| x = 6  y = 3  print(x % y)  print(x // y)  print(x \*\* y)  [결과]  0  2  216 |

### 비교 연산자Comparison Operators

비교 연산자를 통해서 알고자 하는 것은 비교의 결괏값이 True인지, False인지 구분하기 위함이다. 우선 비교 연산자는 크게 6개가 존재 한다. 아래 표를 통해서 확인하도록 한다.

Table 3. Comparison Operators

|  |  |
| --- | --- |
| Operator | Meaning |
| A == B | A와 B가 같다 |
| A != B | A와 B가 같지 않다 |
| A > B | A가 B보다 크다 |
| A < B | A가 B보다 작다 |
| A >= B | A가 B보다 크거나 같다 |
| A <= B | A가 B보다 작거나 같다 |

Python 코드의 예시를 통해 살펴보도록 한다.

|  |
| --- |
| x = 10  y = 11  print(x == y)  print(x != y)  print(x > y)  print(x < y)  print(x >= y)  print(x <= y)  [결과]  False  True  False  True  False  True |

### 논리 연산자Logical Operators

논리 연산자는 주어진 조건식이 True인지 False인지 확인하는데 사용하며, 의사 결정 시에 유용하게 사용된다.

|  |
| --- |
| a = 3  b = 5  print((a > 2) and (b >= 6)  [결과]  False |

위 코드는 Python에서 논리 연산자를 사용하는 예제이다. 논리 연산자는 크게 and, or, not 3개가 존재한다. 논리 연산자의 기본적인 개념을 표로 정리하였다.

Table 4. Logical Operators[[3]](#footnote-3)

|  |  |  |
| --- | --- | --- |
| Operator | Example | Meaning |
| and | X and Y | X와 Y가 모두 True이면 True |
| or | X or Y | X와 Y중 한 개라도 True이면 True |
| not | not X | X가 False이면 True |

다음은 위 논리 연산자의 모든 결과를 보여주는 표이다.

Table 5. Truth Table

|  |  |  |  |
| --- | --- | --- | --- |
| X | Y | X and B | A or B |
| True | True | True | True |
| True | False | False | True |
| False | True | False | True |
| False | False | False | False |

위 표가 실제로 Python에서도 구현이 되는지 코드로 확인한다. 먼저 and 연산자를 구현한 코드이며, 결과는 양쪽이 모두 True일 때, True를 반환한다.

|  |
| --- |
| print(True and True)  print(True and False)  print(False and True)  print(False and False)  [결과]  True  False  False  False |

이번에는 or 연산자를 구현한 코드이며, 결과는 양쪽 중 하나만 True여도 True로 결과가 반환이 된다.

|  |
| --- |
| print(True or True)  print(True or False)  print(False or True)  print(False or False)  [결과]  True  True  True  False |

not 연산자를 사용하면 반대로 결괏값이 반환되는 것도 같이 확인한다.

|  |
| --- |
| print(not True)  print(not False)  [결과]  False  True |

### String

앞에서 String에 대해서 간단하게 다뤘지만, 일반적으로 문자열은 보다 자세히 다뤄야 할 정도로 매우 중요하다.

#### String Operators

우선 문자열에는 덧셈 연산자와 곱셈 연산자 사용이 가능하다. 덧셈 연산자는 두개의 문자를 이어서 내보내고, 곱셈 연산자는 숫자만큼 문자열이 반복된다. 코드를 통해 확인해본다. 먼저 덧셈 연산자는 아래와 같이 표현할 수 있다.

|  |
| --- |
| # 알아서 바꿔요 메롱!  a = "evan"  b = "loves"  c = "sarah"  letter = a + " " + b + " " + c + " forever"  print(letter)  [결과]  evan loves sarah forever |

이번에는 곱셈 연산자를 확인하는 코드를 작성한다.

|  |
| --- |
| x = "Love "  lyrics = x \* 5  print(lyrics)  [결과]  Love Love Love Love Love |

#### 인덱싱Indexing

인덱싱은 각각의 문자열 안에서 범위를 지정하는 것을 의미한다. 다음 이미지[[4]](#footnote-4)를 확인하면 문자열은 앞쪽에서는 0부터 시작하고, 뒤쪽에서는 -1부터 시작하는 것을 확인할 수 있다. 이 때, 기억해야 하는 것은 이번에 배울 인덱싱은 문자열, List, Tuple, array와 같은 자료형에도 동일하게 사용할 수 있다.



Figure 1. String Indexing

위 이미지를 토대로 간단하게 Streamlit Dashboard 문자열을 가지고 인덱싱을 진행하도록 한다 (독자도 임의대로 코드를 작성하여 실행하도록 해본다). 임의의 텍스트를 X 변수에 담고, X[인덱스번호]과 같은 형태로 코드를 실행하면 해당 인덱스번호의 글자가 조회된다.

|  |
| --- |
| X = "Streamlit Dashboard"  print(X[0]) # 0번째 글자 가져오기  print(X[5]) # 5번째 글자 가져오기  print(X[-1]) # -1번째 글자 가져오기  print(X[-3]) # -3번째 글자 가져오기  [결과]  S  m  d  a |

그런데, 만약 주어진 텍스트의 인덱스 번호를 초과하면 다음과 같은 에러가 발생이 된다.

|  |
| --- |
| print(X[100]) |

도표이(가) 표시된 사진

자동 생성된 설명

위 에러 메시지에서 주의 깊게 살펴봐야 하는 것은 ①『----> 1 print(X[100])』와 ②『**IndexError**: string index out of range』이다. ① 메시지는 어떤 코드에서 에러가 나타났는지를 표시해주는 것이고, ② 메시지는 에러코드의 내용이다. 간단히 요약하면, 1번째 print() 코드에서 **IndexError** 에러가 발생했다고 알려주는 것이다. 즉, X가 담고 있는 텍스트는 인덱스 번호가 100까지는 없기 때문에 조회할 것이 없다는 뜻이기도 하다. X 텍스트가 가지고 있는 인덱스 번호를 재 확인하여 입력하면 에러는 발생이 되지 않을 것이다.

#### 슬라이싱Slicing

인덱싱은 각 문자열의 문자를 한 개씩 추출한다면 슬라이싱은 범위를 지정하여 추출하는 것을 말한다. 즉, 기존 데이터에서 일부의 범위를 지정하여 하위 집합을 추출하는 것을 말한다. 이 슬라이싱의 문법은 일반적으로 [start:end:step]의 일반적인 패턴을 따르게 된다.

* start : 슬라이싱의 첫번째 인덱스이며, 해당 인덱스 번호부터 조회를 시작한다. 만약에, 생략을 하면 인덱스는 0번째부터 시작을 한다.
* end : 슬라이싱의 마지막 인덱스를 말한다. 그런데, 해당 인덱스 번호를 포함되지 않고 그 전까지만 조회가 된다. 만약에, 생략을 하면 해당 문자열의 전체 길이가 기본값으로 설정된다.
* step : 슬라이싱의 각 인덱스 사이에서 건너뛸 인덱스의 수를 말하며, 생략할 경우 기본값은 1이다.

주어진 예제를 확인하도록 한다. 이 때, 해당 인덱스 번호를 바꿔서 진행해서 원하는 대로 결과가 나오는지 확인한다.

|  |
| --- |
| greeting = "Hello Streamlit"  print(greeting[:])  print(greeting[6:])  print(greeting[:6])  print(greeting[3:8])  print(greeting[0:9:2])  [결과]  Hello Streamlit  Streamlit  Hello  lo St  HloSr |

인덱싱과 슬라이싱의 원리는 향후에 나오게 되는 List, Tuple 등 다른 자료형에도 그대로 사용이 되기 때문에, 반드시 기억하도록 한다.

#### 문자열 수정

주어진 문자열을 수정하는 코드를 작성한다. 일반적으로는 다음과 같은 코드 형태로 작성할 수 있다. 주어진 문자열 중 H를 A로 바꾸는 코드를 작성한다. 그런데 다음과 같은 에러가 난다.

|  |
| --- |
| greeting = 'Hello, Streamlit'  greeting[0] = 'A'  print(greeting) |
| **---------------------------------------------------------------------------**  **TypeError** Traceback (most recent call last)  **~\AppData\Local\Temp\ipykernel\_29108\2735535411.py** in <module>  1 greeting = **'Hello, Streamlit'**  **----> 2** greeting[0] = **'A'**  3 print(greeting)  **TypeError**: 'str' object does not support item assignment |

우선, 에러가 나는 이유는 문자열은 불변immutable 객체로 인식되기 때문이다. 불변 객체는 객체 생성 이후 내부의 상태가 변하지 않는 것을 말한다. 그렇다면, 새로운 객체를 만들려면 새로운 변수명을 만드는 방법으로 우회할 수 있다.

|  |
| --- |
| greeting = "Hello Streamlit"  new\_greeting = 'A' + greeting[1:]  print(new\_greeting)  [결과]  Aello, Streamlit |

#### String Methods

String 클래스 내부에는 다양한 메서드Methods들이 존재한다.[[5]](#footnote-5) 여기에서 자주 사용하는 여러 메서드들을 확인하도록 한다.

1. str.upper() : 영어 문자들을 모두 대문자로 변경한다.

|  |
| --- |
| sampleText = "hello, streamlit"  upper\_text = sampleText.upper()  print(upper\_text)  [결과]  HELLO, STREAMLIT |

2. str.upper() : 영어 문자들을 모두 소문자로 변경한다.

|  |
| --- |
| sampleText = "HELLO, STREAMLIT"  lower\_text = sampleText.lower()  print(lower\_text)  [결과]  hello, streamlit |

3. str.capitalize() : 영어 문자의 첫번째 글자만 대문자로 변경한다.

|  |
| --- |
| sampleText = "hello, streamlit"  cap\_text = sampleText.capitalize()  print(cap\_text)  [결과]  Hello, streamlit |

4. str.title() : 문자열 내의 각 단어들의 첫번째 글자만 대문자로 변경한다.

|  |
| --- |
| sampleText = "hello, streamlit"  title\_text = sampleText.title()  print(title\_text)  [결과]  Hello, streamlit |

5. str.strip() : 주어진 문자열의 앞뒤 공백이 있다면 모두 제거한다.

|  |
| --- |
| sampleText = " hello, streamlit "  stripped\_text = sampleText.strip()  print(stripped\_text)  [결과]  hello, streamlit |

6. str.replace(old, new) : 기존(old) 문자열을 새로운(new) 문자열로 바꾼다. 주어진 텍스트 hello에서 welcome으로 변경하는 코드를 작성한다.

|  |
| --- |
| sampleText = "hello, streamlit"  new\_text = sampleText.replace("hello", "welcome")  print(new\_text)  [결과]  welcome, streamlit |

7. str.split(sep=None, maxsplit=-1) : 문자열을 sep 파라미터에 입력한 분리 기준점으로 분리한 후 List[[6]](#footnote-6)로 반환한다.

|  |
| --- |
| sampleText = "hello, streamlit"  words = sampleText.split(sep=",")  print(words)  [결과]  ['hello', ' streamlit'] |

8. str.join(iterable) : List나 Tuple내의 값을 하나의 문자열로 변환한다.

|  |
| --- |
| words = ["hello", "streamlit"]  text = ", ".join(words)  print(text)  [결과]  hello, streamlit |

### List

파이썬에 존재하는 자료형으로 시퀀스Sequence 데이터를 다룬다. 시퀀스를 갖는 것은 데이터에 순서가 있다는 뜻이며, 순서가 존재하기 때문에 인덱스와 슬라이싱을 사용할 수 있다. 아래 코드는 주어진 객체가 시퀀스인지 확인하는 코드이다. Sequnece 클래스에 대한 설명은 본 책에서는 생략한다.[[7]](#footnote-7)

|  |
| --- |
| from collections.abc import Sequence  my\_num = 100  my\_list = [1, 2, 3]  my\_string = "hello"  is\_num = isinstance(my\_num, Sequence)  is\_list = isinstance(my\_list, Sequence)  is\_string = isinstance(my\_string, Sequence)  print(is\_num)  print(is\_list)  print(is\_string)  [결과]  False  True  True |

위 결과를 보면 알 수 있듯이 리스트와 문자열은 시퀀스 데이터로 판정을 받았지만, 수치형은 시퀀스 데이터가 아님을 알 수 있다. 그 외 다른 데이터 자료형을 정의하고, 테스트를 해보는 것은 독자에게 맡기도록 한다.

#### 리스트 생성

먼저 리스트를 생성하는 다양한 방법에 대해 살펴보도록 한다. 개발을 하면 종종임의의 빈 리스트Empty List를 생성할 때가 있기 때문에, 관련 코드는 한번씩 숙지하면 좋다. 마지막 변수 e에 저장된 리스트는 리스트 안에 또다른 리스트가 존재하는 것으로 중첩리스트Nested List라고 부르기도 한다.[[8]](#footnote-8)

|  |
| --- |
| a = []  b = list()  c = [2]  d = ['streamlit']  e = [1, 2, ['hello streamlit']]  print(a)  print(b)  print(c)  print(d)  print(e)  [결과]  []  []  [2]  ['streamlit']  [1, 2, ['hello streamlit']] |

#### 인덱싱 & 슬라이싱

이미 문자열에서 인덱싱과 슬라이싱에 관한 설명은 진행했기 때문에 여기에서는 생략하도록 한다.

## pandas

pandas[[9]](#footnote-9)는 데이터 조작 및 분석 도구를 제공하는 Python 라이브러리이다. 일반적으로 CSV 또는 Excel 파일과 같은 구조화된 데이터를 전처리(Processing)하고 다루는(Handling)데 사용된다. Pandas를 사용하면 사용자가 간단하고 효율적인 방식으로 처리하는 불러오고, 정제하고, 필터링하고, 변환하고 시각화를 할 수 있다. 또한 데이터 단순 집계, 그룹화를 활용한 집계, 피벗 테이블에 대한 매우 유용한 기능을 제공한다. 데이터 분석가 뿐만 아니라, 데이터 싸이언티스트, 데이터 엔지니어 등을 포함한 데이터를 다루는 모든 사람에게는 필수적인 도구이다.

일반적으로 pandas는 크게 두개의 객체를 가지고 있다. 먼저 Series 객체는 정수, 실수, 문자열 등 다양한 데이터 유형을 저장할 수 있는 1차원 배열과 유사한 객체이며, Series의 각 요소에 레이블을 지정하는 인덱스가 존재한다. Series를 생성하는 방법은 아래와 같다.

|  |
| --- |
| import pandas as pd  d = {'a': 1, 'b': 2, 'c': 3}  ser = pd.Series(data=d, index=['a', 'b', 'c'])  ser  [결과]  a 1  b 2  c 3  dtype: int64 |

데이터프레임DataFrame 객체는 여러 개의 Series 또는 서로 다른 유형의 배열을 저장할 수 있는 2차원 배열과 유사한 객체로 정의할 수 있다. 각 데이터프레임의 각 요소에 레이블을 지정하는 행, 열 인덱스가 모두 존재한다. 다음은 데이터프레임의 일반적인 예시이다.

|  |
| --- |
| import pandas as pd  d = {'col1': [1, 2], 'col2': [3, 4]}  df = pd.DataFrame(data=d)  print(df)  [결과]  col1 col2  0 1 3  1 2 4 |

데이터 전처리를 하면, 반환값이 Series로 나올 때도 있고, DataFrame으로 나올때도 있기 때문에, 중간에 한번씩 확인하는 것이 좋다.

### 데이터 불러오기

데이터를 불러오는 방법은 pd.read\_csv(각 파일의 경로) 함수를 활용하면 사용할 수 있다. 본 프로젝트와 유사한 부동산 데이터를 불러온다.[[10]](#footnote-10)

1안

|  |
| --- |
| import pandas as pd  df\_boston = pd.read\_csv("./data/boston.csv")  df\_boston |

2안

|  |
| --- |
| import pandas as pd  df\_boston = pd.read\_csv("./data/boston.csv")  df\_boston |

테이블이(가) 표시된 사진

자동 생성된 설명

3안

|  |
| --- |
| import pandas as pd  df\_boston = pd.read\_csv("./data/boston.csv")  print(df\_boston.head())  [결과]  CRIM ZN INDUS CHAS NOX RM AGE DIS RAD TAX \  0 0.00632 18.0 2.31 0.0 0.538 6.575 65.2 4.0900 1.0 296.0  1 0.02731 0.0 7.07 0.0 0.469 6.421 78.9 4.9671 2.0 242.0  2 0.02729 0.0 7.07 0.0 0.469 7.185 61.1 4.9671 2.0 242.0  3 0.03237 0.0 2.18 0.0 0.458 6.998 45.8 6.0622 3.0 222.0  4 0.06905 0.0 2.18 0.0 0.458 7.147 54.2 6.0622 3.0 222.0  PTRATIO B LSTAT target  0 15.3 396.90 4.98 24.0  1 17.8 396.90 9.14 21.6  2 17.8 392.83 4.03 34.7  3 18.7 394.63 2.94 33.4  4 18.7 396.90 5.33 36.2 |

오빠의 선택 :

1. 오빠가 편한 건 사실 1안이 제일 편함.
2. 2안은 사진 캡쳐가 엉성하면 오히려 눈에도 안 들어옴.
3. 3안은 예를 들어서 데이터가 크면 쓸데없는 지면 낭비가 커짐.
4. 출판사에 제출할거면 무조건 2안으로 해야 하나, 우리는 그건 아니니까 약간의 자유도는 존재함.
5. 대충 생각 정리해줘요.
6. 만약 사라가 책을 직접 작성한다면, 어떻게 할 것인가?
7. 사라가 1저자이자 저의 Queen이시니, 전 사라 생각을 따르겠습니다.
8. 일단 전 3안으로 나머지 코드 작업합니다.

### 컬럼 선택

특정한 컬럼을 선택할 때는 List에 특정 컬럼명을 작성하여 아래와 같이 선택할 수 있다. 독자들도 임의의 컬럼을 작성하여 추출하도록 한다.

|  |
| --- |
| cols = ['CRIM', 'ZN', 'INDUS']  result = df\_boston[cols].head()  print(result)  [결과]  CRIM ZN INDUS  0 0.00632 18.0 2.31  1 0.02731 0.0 7.07  2 0.02729 0.0 7.07  3 0.03237 0.0 2.18  4 0.06905 0.0 2.18 |

### 행 선택

행을 선택할 때는 조건식을 통해서 처리하는 경우가 일반적이다. 예를 들어 ZN에서 18.0에 해당하는 행만 선택하는 코드는 아래와 같이 할 수 있다. 조건식을 작성할 때는 비교연산자를 통해서 작성이 가능하다.

|  |
| --- |
| result = df\_boston[df\_boston['ZN'] == 18.0]  cols = ['CRIM', 'ZN', 'INDUS']  print(result[cols])  [결과]  CRIM ZN INDUS  0 0.00632 18.0 2.31 |

### .loc와 iloc

pandas에서 보편적으로 행과 열을 선택하는 방법은 .loc와 .iloc를 사용하는 것이다. 두 메서드의 가장 큰 차이는 loc는 행과 열을 선택할 때, labels/names로 접근하는 방식이고, iloc는 행과 열을 선택할 때, index/position으로 접근하는 방식이다. 만약, loc를 사용해서 행과 열을 선택할 때 매칭되는 label/name이 없다면 key error가 발생한다. 반면, iloc를 사용해서 행과 열을 선택할 때 매칭되는 index가 없다면 index error가 발생한다.

.loc를 활용하여 존재하지 않는 “CRI”를 입력하면 KeyError가 발생한다.

|  |
| --- |
| df\_boston.loc['CRI'] |

텍스트이(가) 표시된 사진

자동 생성된 설명

.iloc를 활용하여 존재하지 않는 컬럼 인덱스 20을 다음과 같이 입력하면 IndexError가 발생한다.

|  |
| --- |
| df\_boston.iloc[:, 20] |

텍스트이(가) 표시된 사진

자동 생성된 설명

#### loc

간단한 예제를 통해 loc가 어떻게 적용되는지 확인한다. 아래 코드는 CRIM, ZN, target 컬럼만 추출하면서, CRIM은 1보다 작은 행을 조회하도록 한다.

|  |
| --- |
| cols = ['CRIM', 'ZN', 'target']  result = df\_boston.loc[df\_boston['CRIM'] < 1, cols]  print(result.head(3))  [결과]  CRIM ZN target  0 0.00632 18.0 24.0  1 0.02731 0.0 21.6  2 0.02729 0.0 34.7 |

이번에는 다중 조건을 입력하도록 한다. 위 조건에 추가적으로 target이 24보다 이상인 데이터만 조회하도록 한다. 아래 코드에서 &은 and 연산자를 의미하고, |은 or 연산자를 의미한다. 독자분들은 & 연산자 대신에 | 연산자를 대입하면 다른 결과가 나오는 것을 확인할 수 있다.

|  |
| --- |
| cols = ['CRIM', 'ZN', 'target']  result = df\_boston.loc[(df\_boston['CRIM'] < 1) &\  (df\_boston['target'] >= 24.0), cols]  print(result.head(3))  [결과]  CRIM ZN target  0 0.00632 18.0 24.0  2 0.02729 0.0 34.7  3 0.03237 0.0 33.4 |

#### iloc

이번에는 iloc를 활용하도록 한다. 일반적으로 iloc를 활용할 때 행은 조건식이 아닌 인덱싱 및 슬라이싱을 활용하여 추출하는 것에 유의한다. 마찬가지로 컬럼도 인덱싱 및 슬라이싱을 활용하여 추출하는 것이다. 예를 들면, 0~4번째 인덱스 행과, 처음 3개의 컬럼만 조회한다고 하면 아래와 같이 작성할 수 있다.

|  |
| --- |
| result = df\_boston.iloc[0:5, 0:3]  print(result)  [결과]  CRIM ZN INDUS  0 0.00632 18.0 2.31  1 0.02731 0.0 7.07  2 0.02729 0.0 7.07  3 0.03237 0.0 2.18  4 0.06905 0.0 2.18 |

만약, 임의의 한 개의 행만 조회한다면, 데이터프레임에서 Series 형태로 바뀐다는 것에 주의한다.

|  |
| --- |
| result = df\_boston.iloc[20, 0:3]  print(result)  print(type(result))  [결과]  CRIM 1.25179  ZN 0.00000  INDUS 8.14000  Name: 20, dtype: float64  <class 'pandas.core.series.Series'> |

### describe()

주어진 데이터의 기초 통계량을 구할 때 쓰는 함수이며, 평균, 표준편차, 각 컬럼의 사분위를 그릴 수 있다.

|  |
| --- |
| cols = ['CRIM', 'ZN', 'target']  print(df\_boston[cols].describe())  [결과]  CRIM ZN target  count 506.000000 506.000000 506.000000  mean 3.613524 11.363636 22.532806  std 8.601545 23.322453 9.197104  min 0.006320 0.000000 5.000000  25% 0.082045 0.000000 17.025000  50% 0.256510 0.000000 21.200000  75% 3.677083 12.500000 25.000000  max 88.976200 100.000000 50.000000 |

## Matplotlib

Streamlit 라이브러리에서는 matplotlib.pyplot figure가 나타나도록 지원하고 있다.[[11]](#footnote-11) 따라서, 기본적인 시각화를 익히도록 한다.

### 선 그래프(Line Chart)

부동산 데이터셋 중 시계열 데이터가 적용된 데이터셋을 불러온 후 간단하게 선 그래프를 그려본다. (데이터셋 참조 : House Property Sales Time Series[[12]](#footnote-12))

|  |
| --- |
| import pandas as pd  import matplotlib.pyplot as plt  sales = pd.read\_csv('data/raw\_sales.csv', parse\_dates=['datesold'])  print(sales.head(3))  [결과]  datesold postcode price propertyType bedrooms  0 2007-02-07 2607 525000 house 4  1 2007-02-27 2906 290000 house 3  2 2007-03-07 2905 328000 house 3 |

데이터셋의 기본 정보는 아래와 같다.

|  |
| --- |
| print(sales.info())  [결과]  <class 'pandas.core.frame.DataFrame'>  RangeIndex: 29580 entries, 0 to 29579  Data columns (total 5 columns):  # Column Non-Null Count Dtype  --- ------ -------------- -----  0 datesold 29580 non-null datetime64[ns]  1 postcode 29580 non-null int64  2 price 29580 non-null int64  3 propertyType 29580 non-null object  4 bedrooms 29580 non-null int64  dtypes: datetime64[ns](1), int64(3), object(1)  memory usage: 1.1+ MB  None |

연도별 평균 주택가격의 선 그래프를 작성하도록 한다.

|  |
| --- |
| import numpy as np  sales['year'] = sales['datesold'].dt.year  result = sales.groupby('year')['price'].agg(np.mean)  fig, ax = plt.subplots(figsize=(10, 6))  ax.plot(result.index, result.values)  ax.set\_title('Avg. House Price per Year')  ax.set\_xlabel('year')  ax.set\_ylabel('price')  plt.savefig('output/matplotlib01.png', dpi=200)  plt.show() |
| 차트이(가) 표시된 사진  자동 생성된 설명 |

## Seaborn

Seaborn 라이브러리를 기본적으로 Matplotlib의 객체를 포함하고 있기 때문에, Streamlit 라이브러리에서 maplotlib 객체로 사용이 가능하다. 따라서, 이번에는 Seaborn 라이브러리로 그래프를 그려보도록 한다.

### 선 그래프(Line Chart)

연도별 평균 주택가격의 선 그래프를 작성하도록 한다.

|  |
| --- |
| import numpy as np  import seaborn as sns  sales['year'] = sales['datesold'].dt.year  result = sales.groupby('year')['price'].agg(np.mean)  fig, ax = plt.subplots(figsize=(10, 6))  sns.lineplot(data=result, x=result.index, y=result.values, ax=ax)  ax.set\_title('Avg. House Price per Year With Seaborn', size = 16)  ax.set\_xlabel('Year', size = 14)  ax.set\_ylabel('Price', size = 14, labelpad=12)  plt.savefig('output/seaborn01.png', dpi=200)  plt.show() |
| 차트이(가) 표시된 사진  자동 생성된 설명 |

## Plotly

Plotly 라이브러리는… .

|  |
| --- |
| import pandas as pd  import plotly  import numpy as np  import plotly.express as px  sales = pd.read\_csv('data/raw\_sales.csv', parse\_dates=['datesold'])  sales['year'] = sales['datesold'].dt.year  result = sales.groupby('year')['price'].agg(np.mean)  fig = px.line(result, x=result.index, y=result.values,  title='Avg. House Price per Year With Plotly')  fig.show() |
| 차트이(가) 표시된 사진  자동 생성된 설명 |

## Scikit-Learn

# Chapter 4. Streamlit

## Slider bar

## Select bar

## Tab

## Checkbox

# Chapter 5. 공공데이터 수집

## 서울열린데이터광장

## 공공데이터포털

# Chapter 6. 부동산 실거래가 대시보드

## 데이터 수집

## Home 화면 꾸미기

## 탐색적 자료분석 화면 꾸미기

## 머신러닝 모형 만들기

## Prediction 화면 꾸미기

## OpenAI 챗봇 구현하기

# Chapter 7. 배포

## secrets.toml

## 배포

## 마무리

# 인용 자료

Jaiswal, S. (2017, Dec). *Python Data Structures Tutorial*. Retrieved from DataCamp: https://www.datacamp.com/tutorial/data-structures-python

1. Jaiswal, S. (2017, Dec). *Python Data Structures Tutorial*. Retrieved from DataCamp: https://www.datacamp.com/tutorial/data-structures-python [↑](#footnote-ref-1)
2. Python 코딩테스트에서는 Stacks, Queues, Graphs, Trees와 같은 개념을 알아야 풀 수 있는 문제들이 있어서, 해당 문법들을 공부할 필요가 있다. 본 책에서는 다루지 않는다. [↑](#footnote-ref-2)
3. 아래 표에서 X와 Y는 모두 조건식을 의미하며, 조건식의 결과가 True 또는 False인 상태를 의미한다. [↑](#footnote-ref-3)
4. https://www.learnbyexample.org/wp-content/uploads/python/String-Indexing.png [↑](#footnote-ref-4)
5. https://docs.python.org/3/library/stdtypes.html#string-methods [↑](#footnote-ref-5)
6. List와 Tuple, Dictionary와 같은 기초 문법은 다음 페이지부터 순차적으로 다룬다. [↑](#footnote-ref-6)
7. 더 읽어보기 : https://docs.python.org/3/library/collections.abc.html [↑](#footnote-ref-7)
8. 일반적으로 리스트에는 파이썬에서 존재하는 다양한 값을 저장할 수 있다. 숫자, 문자열 뿐만 아니라, 아직 배우지 않은 데이터프레임 객체, 함수, 클래스 등도 넣을 수 있어서, 활용범위가 매우 무궁무진하다. [↑](#footnote-ref-8)
9. https://pandas.pydata.org/ [↑](#footnote-ref-9)
10. Boston 데이터셋의 정보는 <http://lib.stat.cmu.edu/datasets/boston> 에서 확인한다. 본 장에서는 독자가 해당 컬럼들의 정보를 이해했다는 가정에서 기술하였다. [↑](#footnote-ref-10)
11. https://docs.streamlit.io/library/api-reference/charts/st.pyplot [↑](#footnote-ref-11)
12. https://www.kaggle.com/datasets/htagholdings/property-sales?resource=download [↑](#footnote-ref-12)