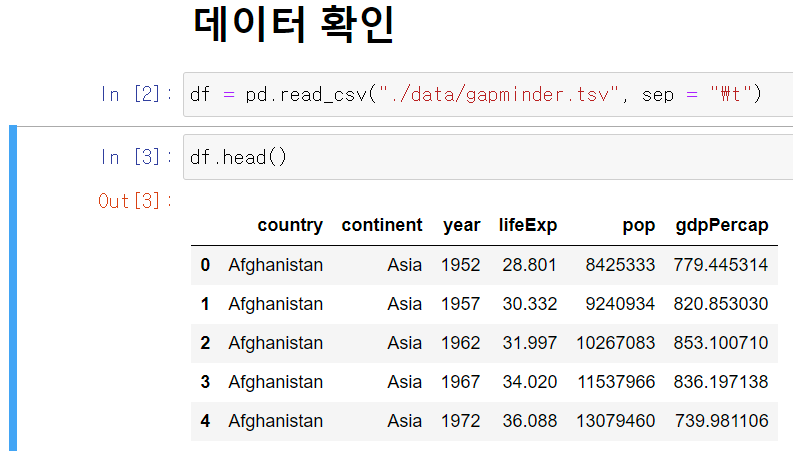
4/25 강의노트

텍스트이(가) 표시된 사진

자동 생성된 설명

데이터프레임 : 판다스 데이터 기본단위 (행과 열로 만들어진 표 형태의 데이터)

판다스(pandas)는 보통 pd라는 별명으로 불러들임



Read\_csv 함수를 통하여 csv 파일을 불러올 수 있음. Csv 함수로 tsv 파일을 불러오기 위해서 구분 단위를 탭으로 바꿔줌 (sep = “\t”)

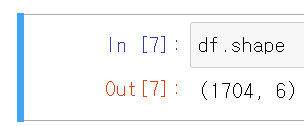
Head 함수를 통하여 불러온 데이터 파일을 출력 가능 (상위 5개 행만)

Csv : 콤마(comma)로 구분된 데이터 파일 형식

Tsv : 탭(Tab)으로 구분된 데이터 파일 형식



타입으로 불러오면 판다스에서 사용되는 데이터프레임 형식으로 인지함



Shape 함수 : 데이터의 크기가 (행수, 열수)의 방식의 튜플로 출력됨

텍스트이(가) 표시된 사진

자동 생성된 설명

행 : 가로로 나열되는 데이터 단위

텍스트이(가) 표시된 사진

자동 생성된 설명

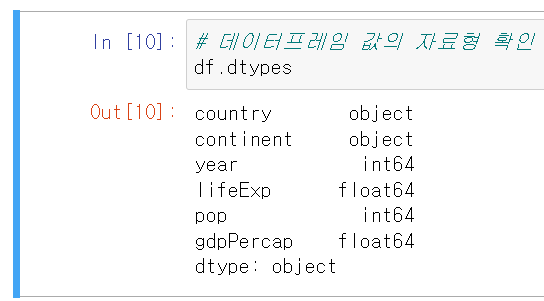
열 : 세로로 나열되는 속성

Column 함수로 각 열의 정보를 출력할 수 있음

텍스트, 영수증이(가) 표시된 사진

자동 생성된 설명

판다스는 데이터가 64비트 형식을 사용하기 때문에 파이썬(32비트)와 데이터 형식이 다르게 나옴



dtypes함수를 이용하면 각 열의 데이터 속성 확인 가능

테이블이(가) 표시된 사진

자동 생성된 설명

info함수 : 데이터의 정보 출력, 위부터

데이터프레임 타입의 데이터

총 1704행의 데이터로 구성됨

각 데이터의 정보 출력 (열 명, NULL값이 아닌 데이터 수, 데이터 타입)

데이터 타입별 개수

메모리 사용량

위와 같이 데이터 정보를 확인하는 함수들은 데이터를 처음 다룰 때 반드시 확인하는 것이 좋음 (head, shapes, dtypes, info 듬)

데이터는 그 범주가 많을수록 유리하기에 행(데이터 개수) 보다는 열(디멘션, 데이터 구분 단위)가 늘어나는 것이 좀 더 유리함

판다스에서는 column명이라는 key 내에 각 행 데이터들에 해당되는 value 값을 보유하는 방식으로 보관함

테이블이(가) 표시된 사진

자동 생성된 설명

이를 series라고 하며 데이터프레임은 series(열)의 집합으로 이루어짐

단일 series는 1개의 열밖에 없으므로 1차원 데이터

데이터프레임은 특정 데이터를 집기 위해서 행과 열의 2가지 정보가 필요하므로 2차원 데이터로 구분됨

테이블이(가) 표시된 사진

자동 생성된 설명

Tail 함수로 데이터의 뒷부분 확인 가능 (데이터프레임도 가능)

테이블이(가) 표시된 사진

자동 생성된 설명

특정 열의 데이터만 추출 가능, 이 경우 각 컬럼 명의 리스트로 나열해야함 (따라서 대괄호가 2개가 됨 하나는 추출 틀, 하나는 리스트를 나타내는 문법)

테이블이(가) 표시된 사진

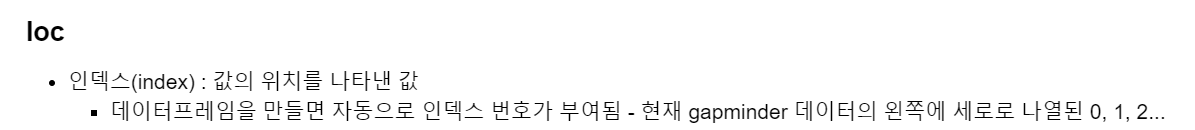
자동 생성된 설명

2개 이상의 열을 추출하면 그것은 2차원이므로 시리즈가 아닌 데이터프레임에 속함

텍스트, 편지이(가) 표시된 사진

자동 생성된 설명

행 단위의 데이터를 추출하는 2가지 방법 존재



인덱스 기반 행 데이터 추출하기

테이블이(가) 표시된 사진

자동 생성된 설명

파란 원 : 인덱스

행 번호는 부여되지 않았을 경우 인덱스와 동일함 (0번 열과 0번째 열은 다를 수 있다)

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

loc으로 데이터를 뽑으면 시리즈(1차원이므로)의 형태로 데이터 출력됨 각 열의 이름이 index로서 오게 됨

Name은 인덱스 넘버

텍스트이(가) 표시된 사진

자동 생성된 설명

데이터프레임 만들기

텍스트이(가) 표시된 사진

자동 생성된 설명

인덱스가 a ~ z 인 데이터프레임이 만들어짐

텍스트이(가) 표시된 사진

자동 생성된 설명

인덱스가 0인 행은 없으므로 loc으로 0을 불러오면 에러가 남 행의 이름을 넣어야 함

텍스트이(가) 표시된 사진

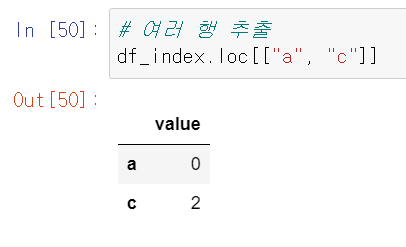
자동 생성된 설명

인덱스가 “a”인 행을 불러옴

텍스트이(가) 표시된 사진

자동 생성된 설명

1개의 행 데이터라도 리스트의 형태로 추출 시 데이터프레임으로 추출이 가능함



여러 행을 데이터프레임으로 추출 가능

텍스트이(가) 표시된 사진

자동 생성된 설명

마지막 행을 loc으로 추출하기 (사실 iloc의 방식임, loc의 방식대로라면 tail 함수 사용하여 마지막 인덱스 확인후 loc 사용)

테이블이(가) 표시된 사진

자동 생성된 설명

Iloc은 열의 순서를 기반으로 데이터를 추출

텍스트이(가) 표시된 사진

자동 생성된 설명

Iloc을 이용하여 마지막 행 추출 (-1은 뒤에서부터 1번째 행)

테이블이(가) 표시된 사진

자동 생성된 설명

여러 행 데이터 추출

테이블이(가) 표시된 사진

자동 생성된 설명

행과 열 추출 loc/iloc [행, 열]

위에서 :는 슬라이싱 문법에 따라 전체 행, 열은 year이므로 전체 year을 추출했음

테이블이(가) 표시된 사진

자동 생성된 설명

2개 이상의 열을 추출할수 있음

행이든 열이든 불러올 때는 행은 행끼리, 열은 열끼리 리스트 형식의 데이터로 만들어야 함

테이블이(가) 표시된 사진

자동 생성된 설명

iloc으로 행과 열 데이터를 추출할 경우에는 열의 순서 또한 0~마지막으로 나열되므로 숫자로 적어야 한다(이름으로 집어넣으면 에러)

테이블이(가) 표시된 사진

자동 생성된 설명

행 열 모두 슬라이싱 형태의 추출

테이블이(가) 표시된 사진

자동 생성된 설명

슬라이싱 문법은 모두 사용 가능

텍스트, 테이블이(가) 표시된 사진

자동 생성된 설명

리스트를 이용한 추출

테이블이(가) 표시된 사진

자동 생성된 설명

describe함수는 수치형 데이터들의 통계 자료 추출

count : 개수, mean : 평균, std : 표준편차, min : 최솟값, 25% : 1사분기값, 50% : 2사분기값, 중간값, 75% : 3사분기값, max : 최댓값

테이블이(가) 표시된 사진

자동 생성된 설명

Groupby 명령어를 통하여 그룹화를 시키고 통계 연산이 가능하다. (year 별로 데이터를 묶어 lifeExp의 평균을 계산함)



여러 개의 기준으로 그룹화하고 여러 개의 열예 대한 평균을 구하는 것도 가능 (리스트 사용)

텍스트, 테이블이(가) 표시된 사진

자동 생성된 설명

그룹화한 데이터 개수 세기

Nunique는 고유한 데이터 개수 세기 = 중복을 제외한 개수

테이블이(가) 표시된 사진

자동 생성된 설명

Value\_counts는 각 고유한 데이터가 각각 몇 개가 있는지 세기

텍스트이(가) 표시된 사진

자동 생성된 설명

Count는 그룹별로 단순 데이터 개수 세기 (NULL은 제외, NULL 포함이면 size 명령어 사용)

텍스트이(가) 표시된 사진

자동 생성된 설명

시리즈 생성, 문자열과 정수타입이 같이 들어감

하나의 데이터는 하나의 데이터 형식을 공유해야 함 “banana”라는 문자열이 들어갔으므로 42가 문자열로 취급됨

실제 데이터를 다룰 때 숫자 사이에 문자가 들어가서 문자열 취급되는 열도 있으니 주의

텍스트이(가) 표시된 사진

자동 생성된 설명

시리즈를 만들 때 index = 를 통하여 인덱스를 직접 부여할 수 있다

텍스트이(가) 표시된 사진

자동 생성된 설명

데이터프레임 생성 : 각 열은 딕셔너리 (key = 열 이름, value = 각 데이터의 집합을 리스트로 표현), 인덱스도 별도 생성 가능하며, columns를 이용하여 나타낼 열도 표현 순서대로 나타낼 수 있음 (이상한거 넣으면 뜨긴 뜨는데 모든 값이 Nan = NULL로 나타남)

테이블이(가) 표시된 사진

자동 생성된 설명

시리즈는 연결된 데이터의 집합이므로 딕셔너리처럼 취급됨

테이블이(가) 표시된 사진

자동 생성된 설명

시리즈에 대해 기본적인 통계 메소드 사용 가능

텍스트, 테이블이(가) 표시된 사진

자동 생성된 설명

데이터 필터링

테이블이(가) 표시된 사진

자동 생성된 설명

조건식을 빼내면 시리즈의 각 열이 모두 조건식에 적용되어 bool 타입으로 출력됨

텍스트이(가) 표시된 사진

자동 생성된 설명

Bool 타입의 리스트를 넣어서 직접 데이터 추출 가능