



DAS 사전교육

- Python 기초 -



Pandas 기초



- ▶ 1. 판다스 개념 및 특징
- ▶ 2. 판다스 객체 생성
- ▶ 3. 판다스 데이터 확인하기
- ▶ 4. 판다스 데이터 선택하기
- ▶ 5. 판다스 결측 데이터 처리하기
- ▶ 6. 판다스 데이터 가공하기
- ▶ 7. 판다스 데이터 그룹핑하기

1. 판다스 개념 및 특징

- ▶ 고수준의 자료구조와 빠르고 쉬운 데이터 분석 도구 제공하는 파이썬 라이브러리
- ▶ 특징
 - ▶ 자동적/명시적으로 축의 이름에 따라 데이터를 정렬할 수 있는 데이터구조
 - ▶ 잘못 정렬된 데이터에 의한 오류를 방지하고, 다양한 방식으로 색인된 데이터를 다룰 수 있는 기능
 - ▶ 통합된 시계열 기능
 - ▶ 시계열 데이터와 비시계열 데이터를 함께 다룰 수 있는 통합 자료구조
 - ▶ 산술 연산과 한 축의 모든 값을 더하는 등 데이터 축약 연산은 축의 이름같은 메타데이터로 전달될 수 있어야 함
 - ▶ 누락된 데이터를 유연하게 처리할 수 있는 기능
 - ▶ SQL 같은 일반 데이터베이스처럼 데이터를 합치고 관계 연산을 수행하는 기능

2. 판다스 객체 생성

- ▶ 시리즈(Series): 레이블을 갖는 1차원 배열
- ▶ 데이터프레임(DataFrame): 레이블을 갖는 행과 열을 갖는 2차원 배열

Series 1			Series 2			Series 3			DataFrame			
Mango			Apple			Banana			Mango Apple Banana			
0	4		0	5		0	2		0	4	5	2
1	5		1	4		1	3		1	5	4	3
2	6	+	2	3	+	2	5	=	2	6	3	5
3	3		3	0		3	2		3	3	0	2
4	1		4	2		4	7		4	1	2	7

3. 판다스 데이터 확인하기

		columns			
		date	temp	max_wind	mean_wind
index	0	2010-08-01	28.7	8.3	3.4
	1	2010-08-02	25.2	8.7	3.8
	2	2010-08-03	22.1	6.3	2.9
	3	2010-08-04	25.3	6.6	4.2
	4	2010-08-05	27.2	9.1	5.6

	3648	2020-07-27	22.1	4.2	1.7
	3649	2020-07-28	21.9	4.5	1.6
	3650	2020-07-29	21.6	3.2	1.0
	3651	2020-07-30	22.9	9.7	2.4
	3652	2020-07-31	25.7	4.8	2.5

3. 판다스 데이터 확인하기

- ▶ 데이터 확인을 위한 함수(메소드)와 속성
 - ▶ 데이터의 (행, 열) 크기 확인 → `df.shape` 속성
 - ▶ 데이터에 대한 전반적인 정보 → `df.info()` 함수
 - ▶ 데이터 앞부분과 마지막 부분 확인 → `df.head()` / `df.tail()` 함수
 - ▶ 인덱스(행 이름)와 열의 레이블(칼럼 이름) → `df.index` / `df.columns` 속성
 - ▶ 데이터의 칼럼별 요약 통계량 → `df.describe()` 함수
 - ▶ 데이터 크기 순으로 정렬 → `df.sort_values()` 함수
 - ▶ 범주형 변수의 빈도분석 결과 → `df.value_counts()` 함수
 - ▶ 열의 고유값 확인 → `df.unique()` 함수



4. 판다스 데이터 선택하기



- ▶ 열 선택하기
- ▶ 행 선택하기
- ▶ 레이블로 선택하기 `df.loc`
- ▶ 위치로 선택하기 `df.iloc`
- ▶ 불 인덱싱

5. 판다스 결측 데이터 처리

- ▶ 결측 데이터 확인
 - ▶ 결측 데이터 개수 확인 → `df.isnull().sum()`
- ▶ 결측 데이터 삭제하기 → `df.dropna(axis, how, thresh, subset, inplace)`
 - ▶ `axis`: 축을 행 또는 열로 결정, 0은 누락된 값이 포함된 행 삭제, 1은 누락된 값이 포함된 열 삭제, 기본값은 0
 - ▶ `how`: `any`는 누락된 값이 있는 경우 행 또는 열을 삭제, `all`은 모든값이 누락된 행 또는 열을 삭제, 기본값은 `any`
- ▶ 결측 데이터 대체하기

6. 판다스 데이터 가공하기

- ▶ 칼럼(변수) 삭제/생성하기
- ▶ 칼럼 이름 변경
 - ▶ `df.columns = ['새변수명1', '새변수명2']` → 전체 변수 이름 재설정
 - ▶ `df.rename(columns = {'기존변수명' : '새변수명'}, inplace = True)` → 원하는 변수 이름만 수정 가능, 딕셔너리 구조로 정의
- ▶ 데이터 형변환

6. 판다스 데이터 가공하기

▶ 데이터 병합하기

- ▶ `pd.merge(df_left, df_right, how = 'inner', on = None)`
- ▶ 아무 옵션을 적용하지 않으면, `on = None` 이므로 두 데이터의 공통 열이름을 기준으로 `inner`(교집합) 조인을 하게 된다.
- ▶ `how = 'outer'` 옵션을 주게 되면, 공통 열이름 기준으로 합치되, 어느 한쪽에라도 없는 데이터가 있는 경우 `NaN`값이 지정된다.
- ▶ 왼쪽에 입력한 데이터프레임 기준(`how = 'left'`)으로, 각각의 `key`값에 해당하는 열을 지정할 수 있다. 기존 데이터프레임에 새로운 데이터프레임으로 파생변수 추가 시 주로 사용된다.



7. 판다스 데이터 그룹핑



- ▶ 한 열을 기준으로 그룹화하기
- ▶ 여러 열을 기준으로 그룹화하기
- ▶ 통계량 관련 함수
 - ▶ sum
 - ▶ mean
 - ▶ std
 - ▶ var
 - ▶ max
 - ▶ min
 - ▶ mode