7. Python

7.1 Pandas 개요

- 1차원 구조를 갖는 시리즈(Series)와 2차원 구조를 갖는 데이터프레임(DataFrame)을 제공
- 데이터프레임은 테이블 형식으로 이종 모음들로 구조화된 데이터로 엑셀 시트 또는 스프레이드시트 형식의 데이터 구조
- 시리즈(Series)는 시계열 데이터를 표현하기 위한 데이터 구조.
  * 시리즈는 데이터프레임의 열(Colume) 한 개를 의미
- 판다스의 데이터프레임과 시리즈 데이터 구조는 재무, 통계, 사회 과학 등 다양한 분야의 데이터 처리에 용이
- 판다스는 데이터의 부분집합 조회, 열 추가 및 제거, 병합, 데이터구조 변경 등 데이터 전처리를 위한 많은 기능을 제공

7.2 Pandas의 장점

- 결측치(Missing Value)처리에 용이
- 데이터 프레임 및 상위 차원 객체에서 열을 손쉽게 삽입, 제거 가능
- 개체를 레이블세트(lableset)에서 명시적인 정렬 및 시리즈, 데이터프레임 등으로 자동 데이터 정렬에 사용
- 데이터를 집계 및 변환을 위해 데이터 세트 분할 및 병합 작업을 수행할 수 있는 유연한 그룹별 기능을 제공
- 다른 파이썬 및 넘파이 데이터 구조의 비정형 색인 데이터를 데이터프레임 형태로 손쉽게 변환
- 지능형 레이블 기반 슬라이싱, 고급 인덱싱 및 대용량 데이터 세트의 하위 집합을 사용 가능
- 데이터 세트의 피벗 및 언피벗 기능을 제공
- 축의 계층적 레이블링(다중 레이블)을 제공
- CSV 파일 또는 구분자에 의한 플랫파일, Excel 파일, 데이터베이스 및 초고속 HDF5 형식의

데이터 저장, 로드를 위한 데이터 입출력 도구를 제공

- 날짜 범위 생성 및 빈도 변환, 선형 회귀 등을 사용 가능

7.3 기초통계 분석에서 사용되는 Pandas 함수

함수	설명
Count	NA를 제외한 개수
Min	최소값
Max	최대값
Sum	합
Cumprod	누적합
Mean	평균
Median	중앙값
Quantile	분위수
Corr	상관관계
Var	표본분산
Std	표본 정규분산