DAS 사전교육

- Python 기초 -

데이터 전처리 기초

- 1. 데이터 전처리 이해
- 2. 데이터 전처리 실습
- 3. 시계열 데이터 전처리
- 4. 시계열 데이터 실습

1. 데이터 전처리 이해

- 데이터 전처리가 필요한 이유
  - 여러 데이터 소스 활용 ➔ 통합 또는 변환 필요
  - 데이터 결측치 또는 오류
  - 이상치
  - 분석 목적에 맞지 않는 변수

1. 데이터 전처리 이해

- 데이터 전처리 방법
  - 데이터 타입의 일관성
  - 결측값 제거 또는 대체
  - 동일한 칼럼값 대체
  - 이상치
  - 목적에 맞는 변수 추출

2. 데이터 전처리 실습

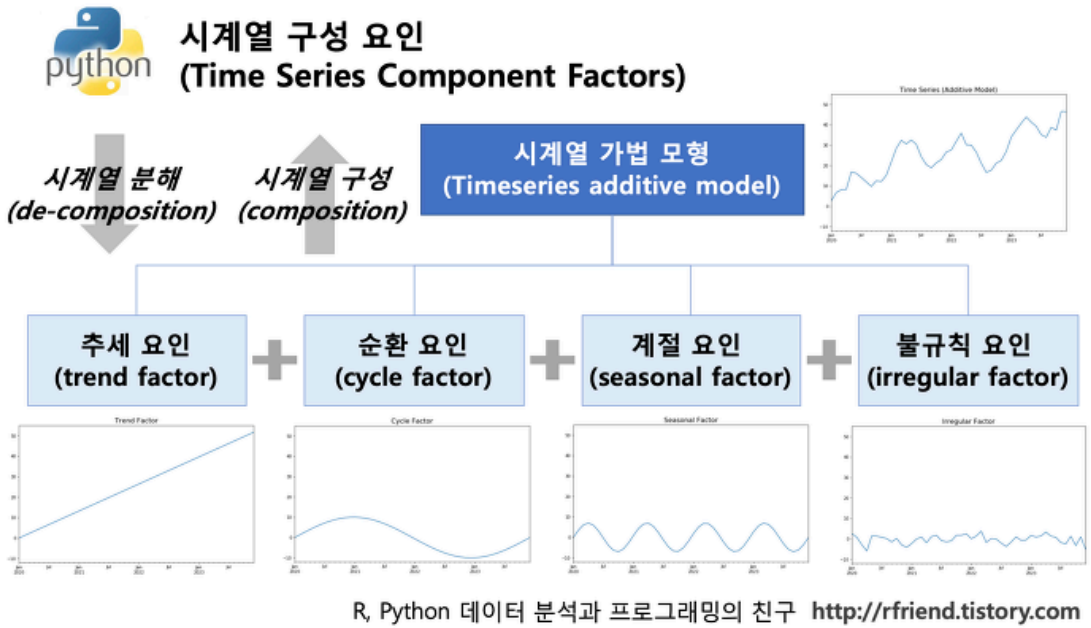
- 데이터

https://www.kaggle.com/datasets/joniarroba/noshowappointments?resource=download

- 문제 정의와 변수
- 데이터 읽기와 확인
- 결손값 확인하기
- 통계량 확인하기
- 데이터 타입 변환
- 새로운 변수 추가
- 값 확인
- 목적에 적합한 변수 추출
- 특성 파악

3. 시계열 데이터 전처리

- 시계열 데이터 패턴
  - 1) 추세(Trend)
  - 2) 주기성(Cycle)
  - 3) 계절성(Seasonality)
  - 4) 불규칙성(Irregular)



3. 시계열 데이터 전처리

- 시계열 데이터 분석
  - 1) AR모형
  - 2) MA모형
  - 3) ARMA모형
  - 4) ARIMA(AutoRegressive Integrated Moving Average)

3. 시계열 데이터 전처리

- 시계열 데이터 전처리 방법
  - 1) datetime 변환
    - https://docs.python.org/3/library/datetime.html#strftime-and-strptime-behavior
  - 2) datetime형의 칼럼을 인덱스로 설정
  - 3) 결측치
  - 4) 빈도 설정
  - 5) 특징량 만들기
  - 6) 이전 값과 차이 계산
  - 7) 카테고리형 컬럼 생성
  - 8) 지연값 추출
  - 9) 원-핫 인코딩

3. 시계열 데이터 전처리

- 4) 빈도 설정
  - asfreq() 옵션

옵션	설명	옵션	설명
B	공휴일과 주말을 제외한 매일(day)	QS	매 분기 시작일
		BQS	공휴일과 주말을 제외한 매 분기 시작일
D	매일(day)	A,Y	매년 마지막일
W	매주 일요일	BA,BY	공휴일과 주말을 제외한 매년 마지막일
M	매월 마지막일	AS,YS	매년 시작일
SM	매월 15일	BAS,BYS	공휴일과 주말을 제외한 매년 시작일
CBM	주말을 제외한 매월 마지막일	BH	공휴일과 주말을 제외한 시(hour)
MS	매월 시작일	H	매시간(hour)
SMS	매월 1일과 15일	T,min	매분(minute)
BMS	공휴일과 주말을 제외한 매월 시작일	S	매초(second)
CBMS	주말을 제외한 매월 시작일	L,ms	매밀리초
Q	매분기 마지막일	U,us	매마이크로초
BQ	공휴일과 주말을 제외한 매 분기 마지막일	N	매나노초

3. 시계열 데이터 전처리

- 5) 특징량 만들기

인덱스	0	1	2	3	4
rolling(1)	5	4	3	2	7
rolling(2)	Null	5	4	3	2
결과	Null	(4+5)/2(rolling수)	(3+4)/2	(2+3)/2	(7+2)/2

4. 시계열 데이터 실습

- 데이터

https://www.kaggle.com/datasets/meetnagadia/apple-stock-price-from-19802021

- 문제 정의와 변수
- 시계열 데이터 확인
- 시계열 데이터 전처리
  - datetime형 변환
  - 날짜로 인덱스 설정
  - 인덱스 중복 여부 확인
  - 다운샘플링
  - 수익률 변수 추가
  - 주가 흐름 파악
- ARIMA 모델