교육 자료

bit.ly/POSCO0511

이장래 강사 Github

<https://github.com/Jangrae>

추가 학습 참고

Do it! 데이터 분석을 위한 판다스 입문

학습 방법

CP0 [참고] 데이터 전처리 기술.ipynb 파일의 코드를 모두 지우고 직접 채우기, 매주???

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

**3. Pandas**

**01 시리즈 만들고 조회**

pd.Series(data, index=[indice])

**02 데이터프레임 만들**기

데이터프레임 만들기

pd.DataFrame(data, index=[indice], columns=[columns])

dict key → column name

pd.read\_csv

난수 발생

np.random.rand()

np.random.randn()

np.random.randint()

날짜 리스트 만들기

pd.date\_range(‘2020-03-01’, periods=10)

인텍스 설정

df.set\_index(‘col01’, inplace=True)

df.index.name = None

인덱스 초기화

df.reset\_index(drop=False, inplace=True)

**03 데이터프레임 정보 확인**

정보확인

df.head()

df.tail()

df.shape

df.index

df.columns

df.values

df.dtypes

df.info()

df.describe() ←- 4분위

인덱스 정렬

df.sort\_index(ascending=True)

컬럼 정렬

df.sort\_value(by= ‘col01’, ascending=False)

df.sort\_value(by= [‘col01’, ‘col02’], ascending=[False, True])

**04 데이터프레임 조회**

df.loc ←- 이름 조회

df.iloc ←- 번호 조회

df[‘col01’, ‘col02’].sort\_values(‘col01’, ascending=True)

df.index.str.contains(‘xxx’)

조건 조회

df.loc[조건] ←- 조건 예) df[‘col01’] > 0.1

isin

between

df[‘col01’].unique()

df[‘col01’].value\_counts(dropna=False)

기본통계

df.sum()

df.mean()

df.max()

df.min()

**05 데이터프레임 변경 1단계**

df[‘col01’].str.replace(‘a’, ‘b’)

np.nan

map({‘a’:1, ‘b’:2, ‘c’:3}) ←- 범주형 데이터 변환

컬럼 이름 변경

df.columns = [‘a’, ‘b’, ‘c’]

df.rename(columns={‘col01’: ‘col\_new’}, inplace=True)

데이터 타입 변경

df[[‘col01’, ‘col02’]] = df[[‘col01’, ‘col02’]].astype(int)

컬럼 삭제

df.drop(‘col01’, axis=1, inplace=True)

**06. 데이터프레임 변경 2단계**

결측치 확인

df[‘col01’].value\_counts(dropna=False)

df.isnull()

df.notnull()

결측치 제거

df.dropna(inplace=True)

df.dropna(subset=[‘col01’], inplace=True)

결측치 채우기

df[‘col01’].fillna(‘xxx’, inplace=True)

직전값 채우기

df[‘col01’].fillna(method= ‘ffill’, inplace=True)

직후값 채우기

df[‘col01’].fillna(method= ‘bfill’, inplace=True)

중복 데이터 확인

df.duplicated(keep= ‘first’)

df.duplicated(subset= ‘col01’, keep= ‘last’)

중복 데이터 제거

df.drop\_duplicates(keep= ‘first’, inplace=True)

가변수 만들기

pd.get\_dummies(df[‘col01’])

pd.concat()

df.drop()

df.rename()

데이터 재 구조화

pd.melt(df, id\_vars= ‘col01’)

pd.melt(df, id\_vars= ‘col01’, var\_name= ‘name01’, value\_name= ‘count’)

Pivot 수행

df.pivot(index= ‘col01’, columns= ‘col02’, values= ‘xxx’)

**07 데이터프레임 집계**

집계

df.groupby(by=[’col01’, ‘col02’], as\_index=False)[‘col03’, ‘col04’].mean()

데이터프레임 합치기

컬럼 합치기

pd.concat((df01, df02), axis=1)

pd.concat((df01, df02[‘col01’]), axis=1)

**08 데이터프레임 조인**

조인

how → inner, outer, left, right

pd.merge(df01, df02, on=’col01’, how=’inner’)

pd.merge(df01, df02, on=[’col01’, ‘col02’], how=’inner’)

**09 데이터프레임 롤링 시프트**

df[‘col\_new’] = df[‘col01’].rolling(3).mean().shift(1)