**PART 3. 데이터 살펴보기**

1. 데이터 프레임의 구조

1-1. 데이터 내용 미리보기

앞부분: df.head(n)

뒷부분: df.tail(n)

디폴트 n=5

1-2. 데이터 요약 정보 확인

데이터프레임의 크기(행, 열)

데이터 프레임의 크기 확인: df.shape

행, 열의 개수를 튜플로 반환

데이터프레임의 기본 정보: df.info()

클래스 유형, 행 인덱스 구성, 열 이름의 종류와 개수, 각 열의 자료형과 개수, 메모리 할당량

데이터프레임의 기술 통계 정보 요약: df.describe()

평균, 표준편차, 최댓값, 최솟값, 중간값

산술데이터가 아닌 열에 대한 정보 포함하려면

Include=’all’ 인자로 추가

고유값 개수, 최빈값, 빈도수 정보 추가

이 경우 산술데이터 가진 열은 NaN

1-3. 데이터 개수 확인

각 열의 데이터 개수

열 데이터 개수 확인: df.count()

시리즈 객체 반환, 유효한 값의 개수만을 계산

각 열의 고유값 개수

열 데이터의 고유값 개수: df[‘열이름’].value\_counts()

Dropna=True 데이터 중 NaN 제외하고 개수 계산

Dropna=False 데이터 중 NaN 포함

2. 통계함수 적용

2-1. 평균값

모든 열(각각의 열)의 평균값: df.mean()

특정 열의 평균값:df[“열이름”].mean()

시리즈로 반환

2-2. 중간값

모든 열: df.median()

특정 열: df[‘열이름’].median()

2-3. 최대값

모든 열: df.max()

특정 열: df[‘열이름’].max()

데이터가 문자열이 포함될 경우 ASCII 숫자로 변환하여 비교

숫자 크기 비교하고 싶다면 문자를 제거하거나 적절한 숫자로 바꾼 후 숫자형 데이터로 변환

2-4.최소값

모든 열: df.min()

특정 열: df[‘열이름’].min()

2-5. 표준편차

모든 열: df.std()

특정 열: df[‘열이름’].std()

문자열 가지는 열은 계산 하지 않음

2-6. 상관계수

모든 열: df.corr()

특정 열: df[열이름의 리스트].corr()

산술데이터 갖는 모든 열에 2개씩 서로 짝짓고, 상관계수 계산

문자열 가진 열은 포함하지 않음

데이터프레임 반환

3. 판다스 내장 그래프 도구 활용

선그래프

Df.plot()

Plot()에 별도 옵션 없는 경우

Df.columns.map(int) 문자열로 저장된 데이터를 정수형으로 변환

Df의 행 인덱스가 그래프의 x축으로 전달

Df의 열 인덱스를 x축으로 전달: df2=df.T

Df2.plot()

행과 열을 전치하여 다시 그림

막대 그래프

Df.plot(kind=’bar’)

히스토그램

Df.plot(kind=’hist’)

산점도

두 변수 사이의 관계를 나타냄

Df.plot(x=’열이름’, y=’열이름’, kind=’scatter’)

박스 플롯

특정 변수의 데이터 분포와 분산 정도 (데이터가 퍼져 있는 정도)

Df[[‘열이름’, ‘열이름’]].plot(kind=’box’)