**Ⅰ. 단변량 데이터 탐색**

**1.1 범주형 자료의 탐색**

* 범주형 변수(categorical variable)는 범주의 수를 파악하기 위해 ‘데이터셋명.value\_counts()’ 함수를 이용한다.
* 윈도표(pie chart)와 막대도표(bar chart)로 시각화 할 수 있다.

텍스트이(가) 표시된 사진

자동 생성된 설명

**1.2. 연속형 자료의 탐색**

* 숫자형(연속형) 자료(numeric/continuos variable)는 describe() 함수를 이용하여

기술통계량을 확인할 수 있다.

텍스트, 달력이(가) 표시된 사진

자동 생성된 설명

* 평균(mean)과 중위수(50%)가 일치할수록 이상치가 적은 데이터를 의미한다.
* 반대로 평균과 중위수의 차이가 크면, 이는 큰 값이든 작은 값이든 이상치가 있다.
* 추가적으로 왜도(skewness)와 첨도(kurtosis)를 살펴보고 정규분포와 유사한지를 검토하는 것이 좋다.

텍스트이(가) 표시된 사진

자동 생성된 설명

[왜도] ‘데이터셋명’.skew() 결과

* 왜도는 0일 경우, 정규분포와 일치하여 좌우가 대칭된 형태이다.

2보다 크면, 어느 한쪽으로 치우쳐져 있다는 의미이다.

* 왜도가 음수(-)면 오른쪽에 자료가 더 많고, 이상치는 작은 값들로 있다는 것이다.
* 왜도가 양수(+)면 왼쪽에 자료가 더 많고, 이상치는 큰 값들로 있다는 것이다.
* 첨도 역시 0이면 정규분포의 높이와 동일하다는 것을 의미한다.

값이 클수록 중심에 자료가 많이 몰려 있다는 것을 의미한다.

※주요 함수

|  |  |
| --- | --- |
| 함수 | 설명 |
| argmin, argmax | 최소, 최대값을 갖고 있는 색인 위치 반환 |
| idxmin, idxmax | 최소, 최대값 갖고 있는 색인의 값 반환 |
| quantile | 0부터 1까지의 분위수 계산 |
| mean | 평균 |
| median | 중위값 |
| mad | 평균값에서 절대 평균편차 |
| var | 표본 분산 |
| std | 표본 표준편차 |
| skew | 표본 비대칭도 |
| kurt | 표본 첨도 |
| cumsum | 누적 합 |
| cumin, cummax | 누적 최소값, 누적 최대값 |
| comprod | 누적 곱 |
| diff | 1차 산술차(시계열 데이터 사용시 유용) |
| pct\_change | 퍼센트 변화율 계산 |
| corr | 데이터프레임의 모든 변수 간 상관관계 계산하여 반환 |
| cov | 데이터프레임의 모든 변수 간 공분산을 계산하여 반환 |

* 연속형 자료의 형태를 가장 확실하게 확인할 수 있는 방법은 각 변수의 분포를 히스토그램으로 확인해보는 것이다. 그러면 정규성 여부, 이상치 존재여부 등을 시각적으로 쉽게 파악할 수 있다.



bins는 구간너비, figsize는 도표 크기를 지정한 옵션

**Ⅱ. 이변량 데이터 탐색**

* corr() 함수로 두 변수 간의 상관계수를 파악할 수 있다.
* 두 연속형 변수 간의 상관은 상관계수 외에 산점도(scatter plot)을 직업 확인하는 것이 좋다. matplolib.pyplot의 scatter 함수에서 X축 변수와 Y축 변수를 각각 설정하면 산점도를 얻을 수 있다.
* 범주별로 종속변수의 평균 및 기술통계량을 알고싶다면, 범주를 groupby 한다.

**Ⅲ. 이상치 처리**

* 이상치는 반드시 분석 전에 반드시 해결해야하는 데이터다.
* 이상치는 보통 제거하거나 또는 데이터 전체 단위를 작게 변환하여 이상치의 영향을 크게 감소시킨다.
* 이상치는 시각적으로 파악하기 위해 상자수염도표(box-whisker plot)를 사용한다.

**3.1. 변수 이상치 처리**

* IQR(Inter-Quartile Range)를 기준으로 이상치를 제거하는 과정이 있다.
* IQR로 하여 이보다 크거나 작은 값은 제거한다.

텍스트이(가) 표시된 사진

자동 생성된 설명