**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

**Ⅳ. 변수 변환**

* 이상치를 제거하지 않은 원데이터를 변환하는 과정이다.
* 대표적으로 원자료에 log를 취하는 방법(log 변환), 제곱근을 취하는 방법(제곱근 변환)이 있다.
* 쉽게 이상치를 해결할 수 있다는 장점이 있지만, 단위가 변화하기 때문에 결과의 해석이 어렵다는 단점도 있다.

**텍스트이(가) 표시된 사진

자동 생성된 설명**

Log 변환 예시

텍스트이(가) 표시된 사진

자동 생성된 설명

제곱근 변환 예시

**Ⅴ. 결측치 처리**

* 데이터프레임에서 결측치(Missing value)는 ‘NaN’으로 표시된다.

**5.1 결측치 확인**

**가. 전체 및 변수별 결측 확인**

- isnull() : 결측이면 True, 결측이 아니면 False를 반환한다

- notnull() : 결측이면 False, 결측이 아니면 True를 반환한다.

- (데이터셋).isnull().sum() : 변수(컬럼)별로 결측값 개수 확인하는 방법

- (데이터셋)[‘컬럼명’].isnull.sum() : 특정 변수(컬럼)의 결측값 개수 확인하는 방법

- (데이터셋).notnull().sum() : 변수(컬럼)별로 결측 아닌 값의 개수 확인하는 방법

- (데이터셋)[‘컬럼명’].notnull.sum() : 특정 변수(컬럼)의 결측 아닌 값의 개수 확인하는 방법

**나. 행별 결측 확인 및 저장**

- (데이터셋).isnull().sum(1) : 행(row) 단위로 결측값 개수 구하기



행(row) 단위로 결측값 개수 구해서 새변수 생성 예시

- (데이터셋).notnull().sum(1) : 행(row) 단위로 실측값 개수 구하기



행(row) 단위로 실측값 개수 구해서 새변수 생성 예시

- (데이터셋)[‘컬럼명’].notnull.sum() : 특정 변수(컬럼)의 결측 아닌 값의 개수 확인하는 방법

**5.2 결측값 제거 : dropna()**

- 측값이 있는 행 제거 : delete row with missing values

- 결측값이 있는 열 제거 : delete column with missing values

- 결측값이 있는 특정 행 또는 열 제거 : delete specific row or column with missing values

1. **결측값 있는 행(row/case) 제거**

- 행(row)은 축(axis)이 ‘0’ 입니다.

- dropna(axis=0)으로 할 경우, 결측이 있는 행(가로)가 제거됩니다.

1. **결측값 있는 열(column/variable) 제거**

- 열(row)은 축(axis)이 ‘1’ 입니다.

- dropna(axis=1)으로 할 경우, 결측이 있는 열(세로)가 제거됩니다.

- 이 경우, 결측치가 있는 변수가 다 제거됩니다.

1. **결측값 있는 특정 행/열 제거**

- 결측이 있는 특정 행을 제거하려면 (데이터셋)[[‘변수명’]].dropna()를 한다.

- 전체 데이터가 아닌 특정 변수 이름을 지정해주어야 한다.

- dropna()와 dropna(axis=0)은 동일하다.

**5.2 결측값 대체 : fillna()**

- 결측값을 특정 값으로 대체 : replace missing values with scalr value

- 결측값을 변수별 평균으로 대체 : filling missing values with mean value per columns

- 결측값을 다른 변수의 값으로 대체 : filling missing values with another columns values

- 결측값을 그룹 평균으로 대체 : filling missing values with another columns values

1. **특정값으로 대체 : df.fillna(value/string)**

- fillna() 안에 바꾸고자 하는 값이나 문자를 설정한다.

- fillna() 옵션에 method를 ffill 또는 pad로 설정하면 결측값을 앞 방향으로 채운다.

앞 방향이란 해당 변수의 앞의 값으로 대체하는 것을 의미한다.

- fillna() 옵션에 method를 bfill 또는 backfill로 설정하면 결측값을 뒷 방향으로 채운다.

뒷 방향이란 해당 변수의 다음 값으로 대체하는 것을 의미한다.

1. **평균 대체**

- (데이터셋).fillna(데이터셋.mean())

- (데이터셋).where(pd.notnull(데이터셋), 데이터셋.mean(), axis=’columns’)

- 중위수로 대체하고자 할 때는 median으로 바꾼다.

- 최대값이나 최소값으로 대체하고자 할 때는 max, min 등으로 옵션을 설정한다.

- 간혹 자기 변수가 아닌 다른 변수 평균으로 대체하고자 할 경우가 있다. 이 때는 fillna()

옵션에 특정 변수의 평균을 설정하면 된다.

1. **다른 변수 값으로 대체**

텍스트이(가) 표시된 사진

자동 생성된 설명

1. **집단 평균값으로 대체**

**텍스트이(가) 표시된 사진

자동 생성된 설명**

평균으로 대체

**텍스트이(가) 표시된 사진

자동 생성된 설명**

설정한 특정 값으로 대체

- 아래 화면은 결측값이 있을 때 salary는 보간법(interpolate), sales는 평균, roe는 ‘missing’ 문자로 대체하는 예시이다.

텍스트이(가) 표시된 사진

자동 생성된 설명

**Ⅵ. 데이터정제 실전과제**

* 데이터프레임에서 결측치(Missing value)는 ‘NaN’으로 표시된다.
* MinMaxScaler : 정규화 (모든 값을 0~1 사이의 값으로 바꾼다)
* StandardScaler : 표준화
* fit() : 데이터에 머신러닝 모델을 맞추는 함수

1. 학습데이터 세트에서 변환을 위한 기반 설정을 하는 함수
2. 데이터를 학습시키는 메서드

* transform() : fit()을 기준으로 얻은 mean, variance에 맞춰서 변형하는 함수

1. fit()을 통해 세운 기반으로 변형하는 함수
2. 실제로 학습시킨 것을 적용하는 메서드