데이터 변환 p186

강창현 | aaakch0316@gmail.com

단위 환산

- df['kpl'].**round(2)** # '연비' 소수점 아래 둘째자리 반올림

데이터 타입_p188

좀 더 이해를 높이기 위해 데이터 타입에 대한 설명을 해보려 한다. 기본적으로 데이터는 범주형과 수치형으로 나눠진다.

- 범주형(Categorical) : 몇 개의 범주로 나누어진 데이터로 명사형(nominal)과 순서형(ordinal)으로 나눠진다.
    - 명사명(nominal) - 순서를 정할 수 없는 데이터로 '성별', '혈액형' 등으로 분류된 자료이다.
    - 순서형(ordinal) - 개개의 값들이 이산적이고 그들 사이에 순서가 존재하는 데이터로 '키', '나이', '몸무게' 등 으로 기준을 분류하거나 나이를 기준으로 '영아' '유아' '어린이' 청소년' ... 으로 분류할 수 있다.

- 수치형(numerical) : 측정(숫자)이 가능한 데이터로 이산형(discrete)과 연속형(continuous)으로 나뉜다.
    - 이산형(discrete) - 연속적이지 않은 수치 데이터로 출산횟수 등을 의미한다.
    - 연속형(continuous) - '키'나 '몸무게' 같은 연속되는 데이터를 의미한다.
      데이터를 원핫인코딩 하려면 카테고리화 해야한다.

자료형 변환

> 정확한 자료형을 확인하는 중이다.

- `df.head()` => (df.dtypes) `df.info()` // CASE : 데이터가 숫자인데, 문자열(object)로 나타난다.
    - 이상하다? => df['열'].**unique()** *
- replace()

```
df['A'].replace('?', np.nan, inplace=True)      # '?'을 np.nan으로 변경
df.dropna(subset=['A'], axis=0, inplace=True)   # 누락데이터 행을 삭제
df['A'] = df['A'].astype('float')      # 문자열을 실수형으로 변환
```

자료형 변환_범주형

CASE : 1, 2, 3으로 표현된 나라 이름을 국가이름으로 변경한다.

```
df['origin'].unique()  # 조회
df['origin'].dtypes    # 정수형
df['origin'].replace({1:'USA', 2:'EU', 3:'JAPAN'}, inplace=True) # 변경
df['origin'].unique() # 변경 확인
df['origin'].dtypes    # 문자형
```

- astype('category') : 이산형 데이터는 **범주형 데이터** 변환
- astype('str') : 문자열 변환

CASE : 국가이름을 범주형으로 변경

```
# origin 열의 문자열 자료형을 범주형으로 변환
df['origin'] = df['origin'].astype('category')
print(df['origin'].dtypes)

# 범주형을 문자열로 다시 변환
df['origin'] = df['origin'].astype('str')
print(df['origin'].dtypes)
```

범주형 데이터 처리_ 매우 중요 p192

- 구간분할
- 더미 변수

구간분할_p192

> 각 구간을 범주형 이상 변수로 변환하는 과정을 구간 분할(binning)이라고 한다. **cut**

```
df['horsepower'] = df['horsepower'].astype('float')      # 문자열을 실수형으로 변환
# np.histogram 함수로 3개의 bin으로 나누는 경계 값의 리스트 구하기
count, bin_dividers = np.histogram(df['horsepower'], bins=3)
bin_names = ['저출력', '보통출력', '고출력']
# pd.cut 함수로 각 데이터를 3개의 bin에 할당
df['hp_bin'] = pd.cut(x=df['horsepower'],     # 데이터 배열
                      bins=bin_dividers,      # 경계 값 리스트[46.  107    168.    230]]
                      labels=bin_names,       # bin 이름
                      include_lowest=True)    # 각 구간의 낮은 경계값 포함
# hp_bin 열의 범주형 데이터를 더미 변수로 변환
horsepower_dummies = pd.get_dummies(df['hp_bin']) # 관련 내용은 뒤
```

더미변수_195

> 범주형 데이터는 훈련 시 바로 사용이 불가능한 경우가 있다. 컴퓨터가 인식 가능한 입력값으로 변환을 해야한다.(원핫 인코딩)
>
> `pd.get_dummies(df['A'])`

- 고유값을 개수만큼 열이 생성된다.

희소행렬(sparse matrix) : 자연어 처리

실습

> 기초4_문제