



CHAPTER

07

데이터 준비 Data Preparation

컴퓨터소프트웨어공학과
김대영

```
import pandas as pd
import numpy as np
from numpy import nan as NA
import re

def main():
    # 7-2
    data = pd.DataFrame(np.random.randn(1000,4))
    print(data.describe())
    col = data[2]
    obj = col[np.abs(col) > 3]
    print(obj)
    obj = data[(np.abs(data) > 3).any(1)]
    print(obj)

    data[np.abs(data) > 3] = np.sign(data) * 3
    print(data.describe())

    df = pd.DataFrame(np.arange(5 * 4).reshape((5,4)))
    print(df)
    sampler = np.random.permutation(5)
    print(sampler)
    obj = df.take(sampler)
    print(obj)

    obj = df.sample(n=3)
    print(obj)

    choices = pd.Series([5, 7, -1, 6, 4])
    draws = choices.sample(n=10, replace=True)
    print(draws)
```

```
df = pd.DataFrame({'key':['b','b','a','c','a','b'],
'data1':range(6)})
obj = pd.get_dummies(df['key'])
print(obj)

dummies = pd.get_dummies(df['key'], prefix='item')
df_with_dummy = df[['data1']].join(dummies)
print(df_with_dummy)

mnames = ['movie_id', 'title', 'genres']
movies =
pd.read_table('TEST/datasets/movielens/movies.dat', sep='
::', header=None, names=mnames)
print(movies[:10])

all_genres=[]
for x in movies.genres:
    all_genres.extend(x.split('|'))
genres = pd.unique(all_genres)
print(genres)

zero_matrix = np.zeros((len(movies), len(genres)))
dummies = pd.DataFrame(zero_matrix, columns=genres)

gen = movies.genres[0]
obj = gen.split('|')
print(obj)
obj = dummies.columns.get_indexer(gen.split('|'))
print(obj)
```

파이썬 실습 2

```
for i, gen in enumerate(movies.genres):
    indices =
dummies.columns.get_indexer(gen.split('|'))
    dummies.iloc[i, indices] = 1

movies_windic =
movies.join(dummies.add_prefix('Genre_'))
print(movies_windic.iloc[0])

val = 'a, b, test'
obj = val.split(',')
print(obj)

pieces = [x.strip() for x in val.split(',')]
print(pieces)

first, second, third = pieces
obj = first + '::' + second + '::' + third
print(obj)

obj = '::'.join(pieces)
print(obj)

print('test' in val)
print(val.index(','))
print(val.find(':'))

print(val.count(','))
print(val.replace(',', '::'))
print(val.replace(',', ''))

text = 'foo bar\t baz \tqux'
obj = re.split('\s+', text)
print(obj)
```

```
regex = re.compile('\s+')
obj = regex.split(text)
print(obj)
obj = regex.findall(text)
print(obj)

text = """ Dave dave@google.com
Steve steve@gmail.com
Rob rob@gmail.com
Ryan ryan@yahoo.com
"""

pattern = r'[A-Z0-9._%+-]+@[A-Z0-9.-]+\.[A-Z]{2,4}'
regex = re.compile(pattern, re.IGNORECASE)
obj = regex.findall(text)
print(obj)

m = regex.search(text)
print(text[m.start():m.end()])
print(regex.match(text))
print(regex.sub('REDACTED', text))

data = {'Dave':'dave@google.com', 'Steve':
'steve@gmail.com', 'Rob':'rob@gmail.com', 'Wes':np.nan}
data = pd.Series(data)
print(data)
pattern = r'([A-Z0-9._%+-]+)@([A-Z0-9.-]+\.[A-Z]{2,4})'
obj = data.str.findall(pattern, re.IGNORECASE)
print(obj)
matches = data.str.match(pattern, re.IGNORECASE)
print(matches)

if __name__ == '__main__': # main()
    main()
```