머신러닝 HW01 20213064_김종민

1. **먼저 아래와 같은 형태로 문작위 데이터를 생성해주세요.**

In [1]:
```
import pandas as pd
import numpy as np
# 샘플 데이터 생성
np.random.seed(0)
data = {
    '이름' : ['박찬호', '류현진', '문동주', '김서현', '주현상'] * 4,
    '과목' : ['수학', '영어', '과학', '국어'] * 5,
    '점수' : np.random.randint(60, 101, 20),
    '학년' : np.random.randint(1, 4, 20)
}
df = pd.DataFrame(data)
```

1. **아래의 문제들을 풀어주세요. 답은** ipynb **파일을** pdf**로 바꿔서 올려주세요**

문제1: groupby**를 사용하여 각 학생의 평균 점수를 계산하고, 평균 점수가 가장 높은 상위** 3**명의 학생을 골라 출력해주세요.**

문제2: groupby**와** apply**를 사용하여 각 과목별로 학년 간 평균 점수 차이가 가장 큰 과목을 구해보세요.**

문제3: apply**를 사용하여 각 학생의 점수에 대해 다음 기준으로등급을 매기는 새로운 열을 만들어 봅시다:**

- 95점 이상: 'A+'
- 90-94점: 'A'
- 85-89점: 'B+'
- 80-84점: 'B'
- 75-79점: 'C+'
- 70-74점: 'C'
- 70점 미만: 'F'

그리고 각 등급별 학생 수를 계산하세요.

문제4: groupby**와** apply**를 사용하여 각 과목별로 상위** 20% **학생의 점수 평균을 계산하세요.**

문제1

In [2]:
```
# 각 학생의 평균 점수를 계산
df1 = df
df1 = df1.groupby('이름').mean(numeric_only = True).drop('학년', axis = 1)
df1.columns = ['평균']
df1
```

Out[2]:

이름	평균
김서현	80.00
류현진	77.75
문동주	78.75
박찬호	80.50
주현상	82.50

In [3]:
```
# 평균 점수가 가장 높은 상위 3명의 학생을 골라 출력
df1.nlargest(3, '평균')
```

Out[3]:

이름	평균
주현상	82.5
박찬호	80.5
김서현	80.0

문제2

In [4]:
```
def calculate_diff(x):
    return x['점수'].max() - x['점수'].min() # 과목별 학년 간 점수 차이 계산
df_calculate_diff = df.groupby('과목').apply(calculate_diff)
df_calculate_diff.idxmax() # 점수 차이 가장 큰 과목 반환
```

Out[4]:
```
'과학'
```

문제3

In [5]:
```
def grading(x):
    if x >= 95:
        return 'A+'
    elif x >= 90:
        return 'A'
    elif x >= 85:
        return 'B+'
    elif x >= 80:
        return 'B'
    elif x >= 75:
        return 'C+'
    elif x >= 70:
        return 'C'
    else:
        return 'F'
df_grade = df
df_grade['등급'] = df['점수'].apply(grading)
df_grade
```

Out[5]:

	이름	과목	점수	학년	등급
0	박찬호	수학	60	2	F
1	류현진	영어	63	2	F
2	문동주	과학	63	1	F
3	김서현	국어	99	2	A+
4	주현상	수학	69	1	F
5	박찬호	영어	79	1	C+
6	류현진	과학	81	2	B
7	문동주	국어	96	3	A+
8	김서현	수학	83	1	B
9	주현상	영어	66	3	F
10	박찬호	과학	84	1	B
11	류현진	국어	84	2	B
12	문동주	수학	72	2	C
13	김서현	영어	61	3	F
14	주현상	과학	98	1	A+
15	박찬호	국어	99	2	A+
16	류현진	수학	83	2	B
17	문동주	영어	84	2	B
18	김서현	과학	77	1	C+
19	주현상	국어	97	3	A+

In [6]:
```
df_grade['등급'].value_counts().sort_index(ascending = True) # 등급별 학생 수
```

Out[6]:
```
A+    5
B     6
C     1
C+    2
F     6
Name: 등급, dtype: int64
```

문제4

In [7]:
```
df_top = df
def quantile_20(subject):
    quantile20 = subject['점수'].quantile(0.8)
    top_20 = subject[subject['점수'] >= quantile20]
    return top_20['점수'].mean()
avg = df_top.groupby('과목').apply(quantile_20)
avg
```

Out[7]:
```
과목
과학    98.0
국어    99.0
수학    83.0
영어    84.0
dtype: float64
```