데이터분석 Hw1

HW1: 성적 분석 과제

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import random
```

1. Create dataset

random.randint를 안쓰고 normal로 쓴 이유는 성적 분포가 정규분포의 형태를 띄는게 일반적이기 때문이다.

```
np.random.seed(42)
# 랜덤한 정규 분포로 학년별 데이터 만들기
grade1 = np.asarray(np.random.normal(loc=65, scale=7, size=(100,3)), dtype = int)
grade2 = np.asarray(np.random.normal(loc=65, scale=7, size=(100,3)), dtype = int)
grade3 = np.asarray(np.random.normal(loc=65, scale=7, size=(100,3)), dtype = int)

# DataFrame변환
g1 = pd.DataFrame(grade1, columns=['Math','English','Korean'])
g2 = pd.DataFrame(grade2, columns=['Math','English','Korean'])
g3 = pd.DataFrame(grade3, columns=['Math','English','Korean'])

# 각 학년 별 합치기
students = pd.concat([g1,g1,g3], ignore_index=True)

# 학년을 나타내는 행 추가
i = np.array([1,2,3])
grade = np.repeat(i,100)
students.insert(0, 'Grade',grade)

print(students)
```

```
     Grade  Math  English  Korean
0        1    68       64      69
1        1    75       63      63
2        1    76       70      61
3        1    68       61      61
4        1    66       51      52
..     ...   ...      ...     ...
295      3    77       69      62
296      3    69       72      70
297      3    68       72      73
298      3    74       69      63
299      3    66       73      59

[300 rows x 4 columns]
```

2. Save dataset .csv

```
students.to_csv('./data/student.csv', index=False)
```

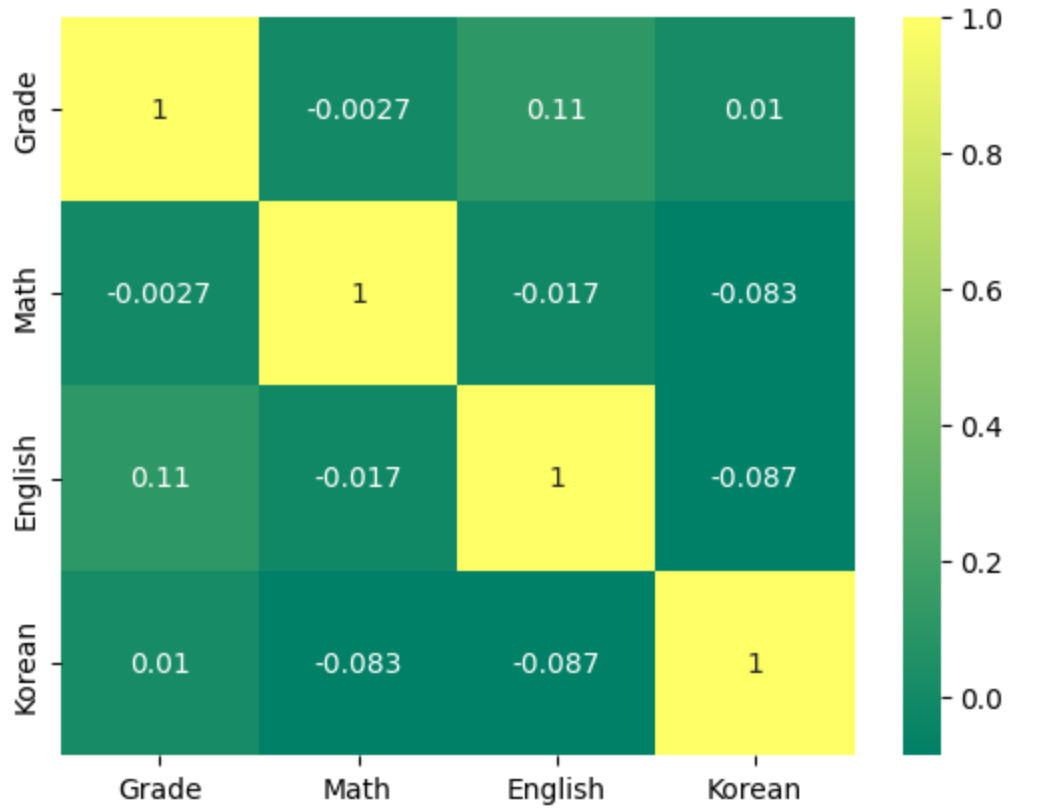
3. Heatmap

```
# import data
data = pd.read_csv('./data/student.csv')

corr_matrix = data.corr()
print(corr_matrix)
```

```
             Grade      Math   English    Korean
Grade     1.000000 -0.002744  0.108850  0.010030
Math     -0.002744  1.000000 -0.017113 -0.082969
English   0.108850 -0.017113  1.000000 -0.087042
Korean    0.010030 -0.082969 -0.087042  1.000000
```

```
# heatmap visualize
import seaborn as sns
sns.heatmap(corr_matrix, annot=True, cmap='summer')
plt.show()
```



해석 : 과목별 상관관계는 없어보인다. 그나마 학년과 영어,한글은 조금 더 관련성을 보인다. 하지만 이도 매우 작은 수치이기 때문에 상관관계를 가지고 영향을 끼친다고 할 수는 없다.

4. Describe()

```
data.describe()
```

	Grade	Math	English	Korean
count	300.000000	300.000000	300.000000	300.000000
mean	2.000000	65.116667	63.820000	65.073333
std	0.817861	5.961110	6.874965	7.746482
min	1.000000	50.000000	42.000000	46.000000
25%	1.000000	61.000000	59.000000	59.000000
50%	2.000000	66.000000	64.000000	65.000000
75%	3.000000	68.000000	69.000000	70.000000
max	3.000000	83.000000	82.000000	91.000000

random 데이터를 만들 때 평균을 65로 잡았기 대문에 평균이 다 같은 것을 볼 수 있다.

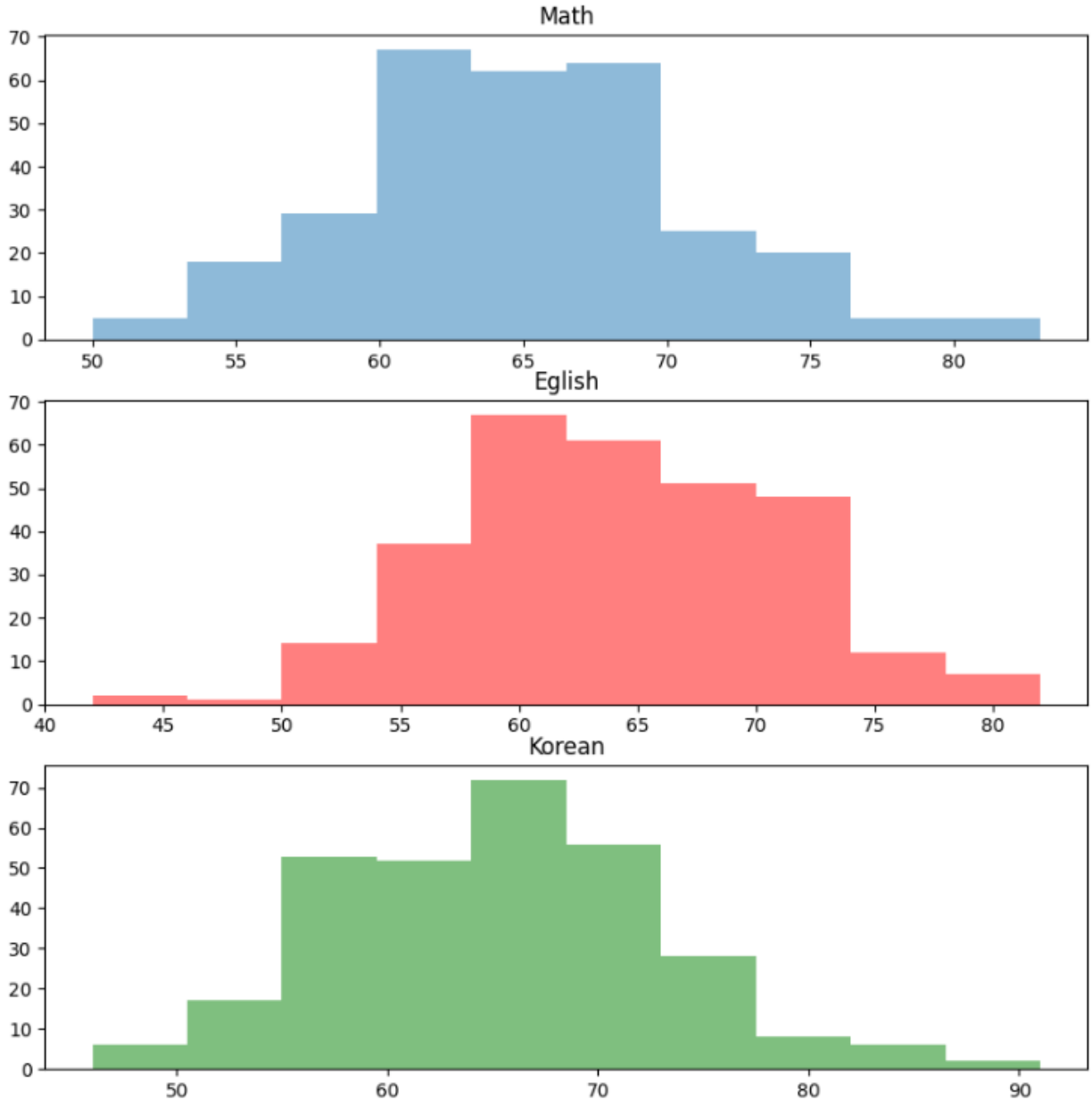
5. Normalize

```
#Normal Distribution
fig = plt.figure(figsize=(10,10))

plt.subplot(311)
plt.hist(data['Math'], bins=10, alpha=0.5)
plt.title('Math')

plt.subplot(312)
plt.hist(data['English'], bins=10, alpha=0.5, color='r')
plt.title('Eglish')

plt.subplot(313)
plt.hist(data['Korean'], bins=10, alpha=0.5, color ='g')
plt.title('Korean')
```


Math
70
60
50
40
30
20
10
0
50
55
60
65
70
75
80
Eglish
70
60
50
40
30
20
10
0
40
45
50
55
60
65
70
75
80
Korean
70
60
50
40
30
20
10
0
50
60
70
80
90