201502119 정지원 lab01 보고서

공통

- 데이터 전처리 내용

```
3   '''
4   행정구역            총 25개.
5   계, 남, 여          총 3개.
6   나이 0 ~ 100+       총 101개.
7   [행정구역][계|남자|여자][나이]
8   '''
9
10  classify_name = ['계','남자','여자']
11  axis1, axis2, axis3 = (25, 3, 101)
12
13  def parseData():
14      with open('seoul.txt', mode='r', encoding='utf-8') as f:
15          category = f.readline() # pass
16          data = np.zeros((axis1, axis2, axis3),np.int64)
17          for idx, line in enumerate(f):
18              arr = line.split('\t')
19              data[idx // axis2, idx % axis2] = np.array(arr[2:])
20          data = np.array(list(sum(data[j,i] for j in range(axis1)) for i in range(axis2)))
21      return data
```

seoul.txt의 첫 줄은 카테고리와 행의 이름이므로 pass 한다. (f.readline())

두 번째 줄부터 3줄씩 **행정구역, [계|남자|여자] (성별), 나이별 데이터 101개** 이고, 서울시 행정구역은 총 25개 이다. 즉, 실질적인 데이터는 두 번째 줄부터 각 줄마다 세 번째부터 시작하기 때문에, line 19에서 arr이 인덱스 2부터 시작하였다.

이를 [행정구역][성별][나이] 로 데이터를 넣은 뒤, 성별이 같은 데이터를 묶어 반환하였다.

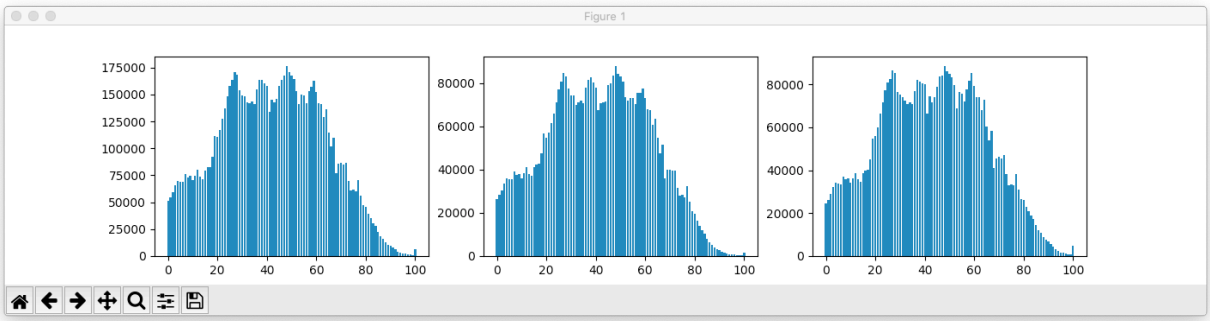
반환된 값은 [list of 행정구역별 `계`의 총합, list of 행정구역별 `남자`의 총합, list of 행정구역별 `여자`의 총합] 이다.

```
data = np.array(list(sum(data[j,i] for j in range(axis1)) for i in
range(axis2)))
```

axis2는 성별에 따른 축이고, axis1은 행정구역에 따른 축이다.

- 분산을 구할 때 4바이트를 넘어가므로 데이터형은 int64로 통일하였다.

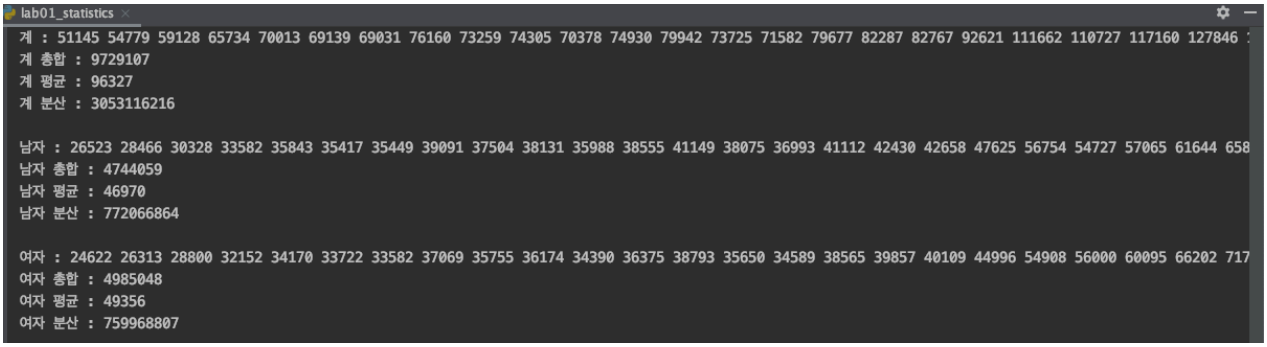
goal 1



x축은 나이(세), y축은 인구 수(명)

xlabel에서 한글이 깨져 영어로 대체하였다.

goal 2



numpy의 mean()과 var()를 사용하여 평균과 함수를 구하였다.