1.(통실_corona.csv)

corona.csv 데이터는 2020년 6월 1일 부터 2020년 8월 31일 까지 국내 코로나 확진자 수를 나이별로 정리해놓은 데이터이다.(시간역순으로 정렬되어있음, 0행이 2020년 8월 31일 확진자이고 91행이 2020년 6월 1일 확진자임) row는 한 날에 새로 확진된 환자 수이고, column은 확진자의 나이대이다.(ex 0이면 만 0-9세, 3이면 만 30-39세, 8은 만80세 이상)

\\ 매일 확진되는 환자수는 독립이라 가정하자.

In [ ]:

```
corona = pd.read_csv('통실_corona.csv')
```

In [ ]:

```
corona
```

위에 적힌 함수를 이용하여 데이터를 load하자.

데이터를 살펴보면 강의록에서 ANOVA 검정할 때 사용하는 모양과는 다른 것을 확인 할 수 있다.

(1).

corona.values.reshape(279)와 np.tile 함수를 이용하여 아래와 같은 데이터프레임을 만들고 6월 부터 8월까지 확진자수는 총 몇명인지 쓰시오.

In [ ]:

```
corona
```

(2).

나이대별 확진자 수의 평균이 같은지를 검정하고자 한다. 귀무가설과 대립가설을 각각 쓰시오.

위 데이터를 이용하여 유의수준 5%이내에서 검정을 하였을 때, anova table을 출력하고, 검정통계량(F값)과 Pvalue와 기각여부를 쓰시오.

(3).

ANOVA 검정은 각 집단의 분포가 등분산이고 정규분포임을 가정한다. 이를 확인하기 위해 QQ-plot을 사용하고자 한다. 하지만 현재 model.resid를 출력하면, 나이대 별로 나뉘어있지 않으므로 나이대 별로 나눈 후 9개의 QQ-plot을 그려야한다. 편의상 등분산은 가정하고, 10대와 20대의 QQ-plot만을 그려보자. model.resid에서 10대와 20대에 해당하는 값들을 모아 따로 (리스트형태로) resid_10과 resid_20으로 저장하고 이를 이용하여 QQ-plot을 그려보고 정규성을 위반하는지 여부를 서술하시오.(주의!! : ProbPlot() 함수에 list를 input하면 오류가 뜨므로 list를 np.array를 이용하여 numpy array로 변환하고 input해야함 ex:ProbPlot(np.array(resid_20)))

2.(run10Samp)

run10Samp.txt 데이터는 2012년 Cherry Blossom 10 mile run 경기에서 완주를 한 선수 100명의 자료이다. 각 column은 완주시간,나이,성별,출신지역이다.

In [ ]:

```
run10Samp = pd.read_csv('run10Samp.txt',sep=" ")
```

(1).

성별에 따라 완주시간에 차이가 있는지를 일원배치법을 이용하여 유의수준 5%에서 검정해보자. 가설을 세우고 검정통계량(F값), P_value를 쓰고 귀무가설 기각여부를 쓰시오.

(2).

QQ-plot을 이용하여 성별에 따른 완주시간의 분포가 정규분포를 따르는지 확인하시오. 아래의 함수는 model.resid에서 gender가 female일 때와 male일 때의 index를 저장한 리스트이다.

In [ ]:

```
female_list = np.where(run10Samp['gender']=='F')[0]+1
male_list=np.where(run10Samp['gender']=='M')[0]+1
```

(3).

등분산을 가정하고 성별에 따른 완주시간에 차이가 있는지를 t-test를 이용하여 유의수준 5%에서 검정해보자. 가설을 세우고 검정통계량값(t값), P_value를 쓰고 귀무가설 기각여부를 쓰시오.

(4).

(1)에서의 검정통계량 값과 (2)에서의 검정통계량 값은 어떠한 관계가 있고, 왜 이러한 관계가 있는지 쓰시오.

(5).

run10Samp의 age를 정확한 나이가 아닌 위의 문제와 같이 연령대로 나누어 볼 것이다. run10Samp의 age열을 나이대로 바꾸시오. (58->5,42->4) \ 나이대와 성별에 따라 완주시간에 차이가 있는지를 반복이 없는 이원배치법을 이용하여 검정하고자 한다. 귀무가설과 대립가설을 쓰시오.

(6).

성별과 연령이 완주시간에 영향을 미치는지 유의수준 5%에서 검정해보자. 검정통계량(F값), P_value를 쓰고 귀무가설 기각여부를 쓰시오.

(7).

성별과 연령에 대한 상호작용 유무를 interaction plot을 통해 확인해보시오