통계계산프로그래밍 기말대체과제

2019150445 통계학과 신백록

Part1

1.1)

```
tab.1 <- sort(tab.0)

barplot(tab.1,horiz=T,main='NYC Flights 2013',las=1)
```

1.2)

```
names(tab.0)[rk %in% 1:10]

flights[flights$carrier %in% carrier.top10,]
```

2.

```
is.na(flights.1$dep_delay)|flights.1$dep_delay>=60


x<- tab.2m / tab.2m[,3]

x.1<-x[c(order(x[-11,1]),11),]

round(x.1,3)
```

3.

```
tab.3m <- round(tab.3m / tab.3m[,3],3)

mosaicplot(tab.3m[,-3],col=c('grey','red'),off=0,main='NYC Flights 2013')
```

Part2

```
library(tidyverse)

library(Lahman)

data(Teams)
```

1)

```
Teams %>% filter(lgID %in% c('AL','NL'),yearID>=1975) %>% group_by(lgID,yearID) %>% summarize(attendance=sum(attendance)) %>% spread(lgID,attendance)
```

```
Teams %>% filter(lgID %in% c('AL','NL'),yearID>=1975) %>% group_by(lgID,yearID) %>% summarize(attendance=sum(attendance)) %>% ggplot(aes(x=yearID, y=attendance/1000000))+geom_line(aes(col=lgID))+ylab('attendance(millions)')+xlab('year')
```

위 plot을 살펴보면, AL과 NL 두 그래프는 꽤나 비슷한 경향을 보이고 있다. 특히 1981년쯤에 관객의 수가 급격히 감소한 점, 그 후에 계속 우상향 그래프를 그리다가 1993~1995년 즈음에 다시 한 번 급격히 관객의 수가 감소한 것이 눈에 띈다. 그 후에는 상승과 하락을 반복하지만 전체적으로 봤을 때는 두 그래프 모두 관객의 수가 증가하고 있다고 보여진다.

2)

```
Teams %>% mutate(year=substr(as.character(Teams$yearID),3,4)) %>% filter(teamID=='BOS',yearID>=1975) %>% ggplot(aes(x=W/G, y=attendance/1000000,
```

label=year))+geom_text(size=3)+xlab('WR')+ylab('Attendance(millions)')

Teams %>% mutate(year=substr(as.character(Teams$yearID),3,4)) %>% filter(teamID=='LAN', yearID>=1975) %>% ggplot(aes(x=W/G, y=attendance/1000000, label=year)) + geom_text(size=3)+xlab('WR')+ylab('Attendance(millions)')

두 그래프를 비교하기 위하여 하나의 grid에 나타내면,

Teams %>% mutate(year=substr(as.character(Teams$yearID),3,4)) %>% filter(teamID %in% c('LAN','BOS'),yearID>=1975) %>% group_by(teamID) %>% ggplot(aes(x=W/G, y=attendance/1000000,label=year))+geom_text(size=3)+xlab('WR')+ylab('Attendance(millions)')+facet_grid(.~teamID)

위 그래프를 살펴보면, Boston Redsox보다 LA Dodgers가 그래프가 오른쪽 위로 치우쳐져 있고, 대체적으로 관중과 승률 모두 Boston보다 높은 것을 볼 수 있다. 또한 1981년도의 Boston Redsox의 관중 수는 심각한 outlier로, 아까 1번에서 보았듯이 1981년의 관중 수가 급감한 것을 이 산점도에서도 확인할 수 있다. LA Dodgers도 1981년도에 다른 년도에 비해 관객수가 적긴 하지만, Boston 등 다른 League의 팀들이 감소한 것에 비해서는 그렇게 크게 감소하지 않은 것을 볼 수 있다.

3)

활동 년 수를 팀에 몇 년 동안 소속되어 있었는 지로 해석할 수도 있을 것이고,

가장 최근에 뛰었던 연도와 데뷔년도의 차이로 해석할 수도 있을 것이다.

중간에 은퇴를 번복했거나, 휴식기를 갖고 다시 팀에 입단한 경우도 있기 때문에 둘의 데이터는 달라진다.

예를 들어 다음과 같은 코딩으로 playerID가 allisdo01인 선수의 활동년도를 보면 다음과 같은 결과가 나온다.

Batting %>% filter(playerID=='allisdo01') %>% select(yearID)

```
yearID
1   1871
```

```
2    1872
3    1872
4    1873
5    1873
6    1874
7    1875
8    1876
9    1877
10   1878
11   1879
12   1883
```

여기서 11행과 12행을 보면 알 수 있듯이, 중간에 4년을 쉬었다가 다시 1883년에 팀에 입단을 하였다. 이 선수가 실제로 활동한 년 수는 10년(71,72,73,74,75,76,77,78,79,83)이고, 휴식기까지 활동 년 수로 친다면 13(1883-1871+1)년이 된다.

두 경우 모두 확인해보자.

활동 년 수=휴식기까지 포함한 데뷔 년도~가장 최근 년도

```
Batting %>% group_by(playerID) %>% summarize(range=diff(range(yearID))+1) %>% ggplot(aes(x=range))+geom_histogram(col='black',bins=35)
```

```
Batting %>% group_by(playerID) %>% summarize(range=diff(range(yearID))+1) %>% arrange(-range) %>% head(1)
```

→'altroni01'이 36년으로 가장 오래 활동

```
Batting %>% filter(playerID=='altroni01') %>% select(yearID,G,AB,H)
```

활동 년 수=실제 팀에 소속되어있는 년 수

```
Batting %>% group_by(playerID) %>% summarize(range=n_distinct(yearID)) %>% ggplot(aes(x=range))+geom_histogram(col='black',bins=25)
```

```
Batting %>% group_by(playerID) %>% summarize(range=n_distinct(yearID)) %>% arrange(-range) %>% head(1)
```

→'ansonca01'이 27년으로 가장 오래 활동

Batting %>% filter(playerID=='ansonca01') %>% select(yearID,G,AB,H)

4)

Batting %>% group_by(playerID) %>% filter(sum(AB)>=5000) %>% summarize(BA=sum(H)/sum(AB)) %>% arrange(-BA) %>% head(1)

→'cobbty01'이 타율 0.366으로 총 타석 수가 5000 이상인 선수 중 가장 높은 커리어 타율을 보임

Batting %>% filter(playerID=='cobbty01') %>% summarize(Career_H=sum(H), Carrer_AB=sum(AB), BA=sum(H)/sum(AB))