통계계산프로그래밍 기말시험 대체과제

마감: 2020년 12월 15일(화) 23:59

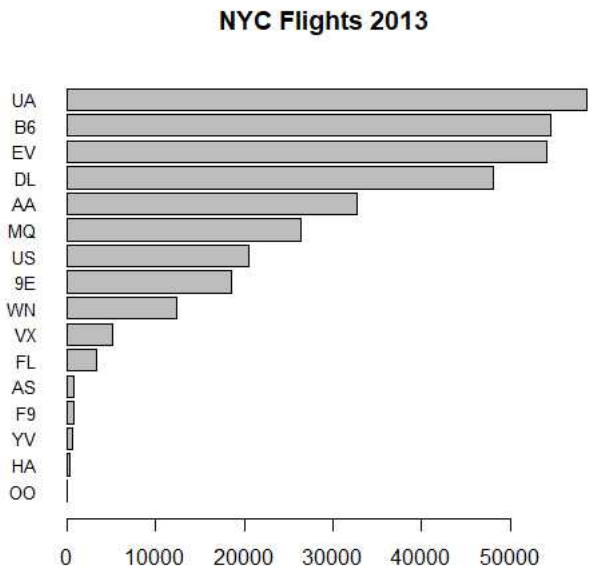
파트2 힌트: 용어 및 분석스킴은 R프로그래밍 부록. R을 활용한 MLB데이터분석 참조

파트1.(20점)

`nycflights13` 데이터셋은 `2013.1.1. ~ 2013.12.31.` 동안의 미국 뉴욕 인근 3개의 국제공항인 케네디(`J.F.Kennedy`)공항(`JFK`), 뉴어크(`Newark`)공항(`EWR`), 라과디아 공항(`LGA`)에서 출발이 계획되어 있던 모든 비행편에 대한 비행 정보를 기록한 데이터이다.
데이터셋에 있는 5개의 테이블 중 `flights` 테이블은 비행편에 대한 정보를 담고 있다.

`nycflights13::flights` 데이터(336,776개 관측, 19개 변수)로부터 다음 질문에 답하라.
(빈칸에 적절한 R코드를 작성하시오)

1. (10점) 1) 항공사 `carrier`별 빈도를 아래와 같은 막대그림으로 나타내라. 밑줄친 곳에 적당한 R표현은?



풀이. R 스크립트

```
library(nycflights13)
data(flights)
str(flights)
tab.0 <- table(flights$carrier)
_____________________________________________
_____________________________________________
```

2) 그 중 상위 10개 항공사만으로 부 데이터 **flights.1**을 구성하여 이후 풀이에 적용하라. **flights.1**이 총 데이터를 커버하는 비율은 몇 퍼센트인가?
밑줄친 곳에 적당한 R표현은?

풀이. R 스크립트

```
rk <- rank(-tab.0); rk
carrier.top10 <- _____________________________________
flights.1 <- _____________________________________
str(flights.1)
round(nrow(flights.1)/nrow(flights)*100, 1)
```

답 (출력). 98.3%

2.(5점) 나쁜 항공편 **badflight**를 출발지연 **dep_delay**가 NA이거나 60분 이상인 경우로 정의하자. 항공사 별 나쁜 항공편의 비율을 다음과 같이 순서 정렬된 표로 제시하라.
밑줄친 곳에 적당한 R표현은?

```
        badflight
carrier  good   bad   Sum
     EV  0.819 0.181 1.000
     9E  0.836 0.164 1.000
     MQ  0.876 0.124 1.000
     WN  0.896 0.104 1.000
     B6  0.906 0.094 1.000
     AA  0.918 0.082 1.000
     UA  0.922 0.078 1.000
     VX  0.923 0.077 1.000
     US  0.930 0.070 1.000
     DL  0.937 0.063 1.000
    Sum 0.895 0.105 1.000
```

풀이. R 스크립트

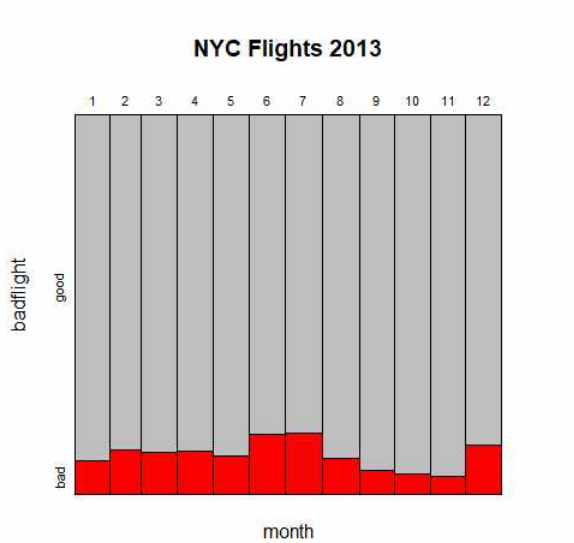
```
flights.1$badflight <-
    as.factor(_____________________________________________________)
levels(flights.1$badflight) <- c("good","bad")
tab.2 <- with(flights.1, table(carrier, badflight))
tab.2m <- addmargins(tab.2)
_____________________________________________________
_____________________________________________________
_____________________________________________________
```

4.(5점) 나쁜 항공편이 월(month)과 관련이 있는가를 아래와 같이 2원표와 모자이크 플롯으로 나타내라. 나쁜 항공편 비율이 높은 3개 달은 6월, 7월, 12월이다.

밑줄친 곳에 적당한 R표현은?

badflight

month	good	bad	Sum
1	0.912	0.088	1.000
2	0.881	0.119	1.000
3	0.887	0.113	1.000
4	0.885	0.115	1.000
5	0.899	0.101	1.000
6	0.840	0.160	1.000
7	0.837	0.163	1.000
8	0.904	0.096	1.000
9	0.935	0.065	1.000
10	0.945	0.055	1.000
11	0.951	0.049	1.000
12	0.871	0.129	1.000
Sum	0.895	0.105	1.000



풀이. R 스크립트

```
tab.3 <- with(flights.1, table(month, badflight))
tab.3m <- addmargins(tab.3)
_____________________________________________________________
round(tab.3m, 3)
mosaicplot(_____________________________________________________________,
    main="NYC Flights 2013")
```

파트2.(20점)

메이저리그(Lahman 팩키지의 Salaries, Batting, Pitching)의 2015년 데이터에 대하여 다음 질문들에 답하라.

1. Lahman 팩키지의 Teams 데이터로부터 1975년 이후 리그(American League, National League) 별 연 관중 수(attendance)의 시도표를 제시하고 주요 특징을 기술하라.

   도움말: `aggregate()`와 `ts()` 활용

2. [앞 문제의 계속] 1975년 이후 Boston Redsox 팀에 대하여 관중 수 (= y축) 대 팀 승률 (= x축)의 산점도를 만들어라 (이때 좌표에 점 대신 연도의 마지막 두 숫자를 넣어라). LA Dodgers 팀에 대하여 같은 질문에 답하라. 두 그래프는 어떻게 다른가?

   도움말: 유사 코딩의 반복을 피하기 위해서 사용자 함수를 만들어 사용할 필요가 있다.

3. Lahman 팩키지의 Batting 데이터로부터 모든 선수의 활동년 수를 산출하여 히스토그램으로 제시하라. 활동년 수가 가장 큰 선수를 찾아 그의 연도별 활동(G, AB, H)을 살펴보라. (위키피디어에서 그를 찾아보라) *G:게임수, AB:타석, H:안타

   도움말: `plyr::ddply()` 함수와 `which.max()` 함수를 활용

4. [앞 문제의 계속] 메이저리그 역사에서 Career 타율(BA, batting average)이 가장 좋은 선수는 누구인가? 단, 총 타석 수를 5,000 이상으로 조건화한다. 그의 Career 안타수, Caree 타석수는 얼마였는가? (위키피디어에서 그를 찾아보라)