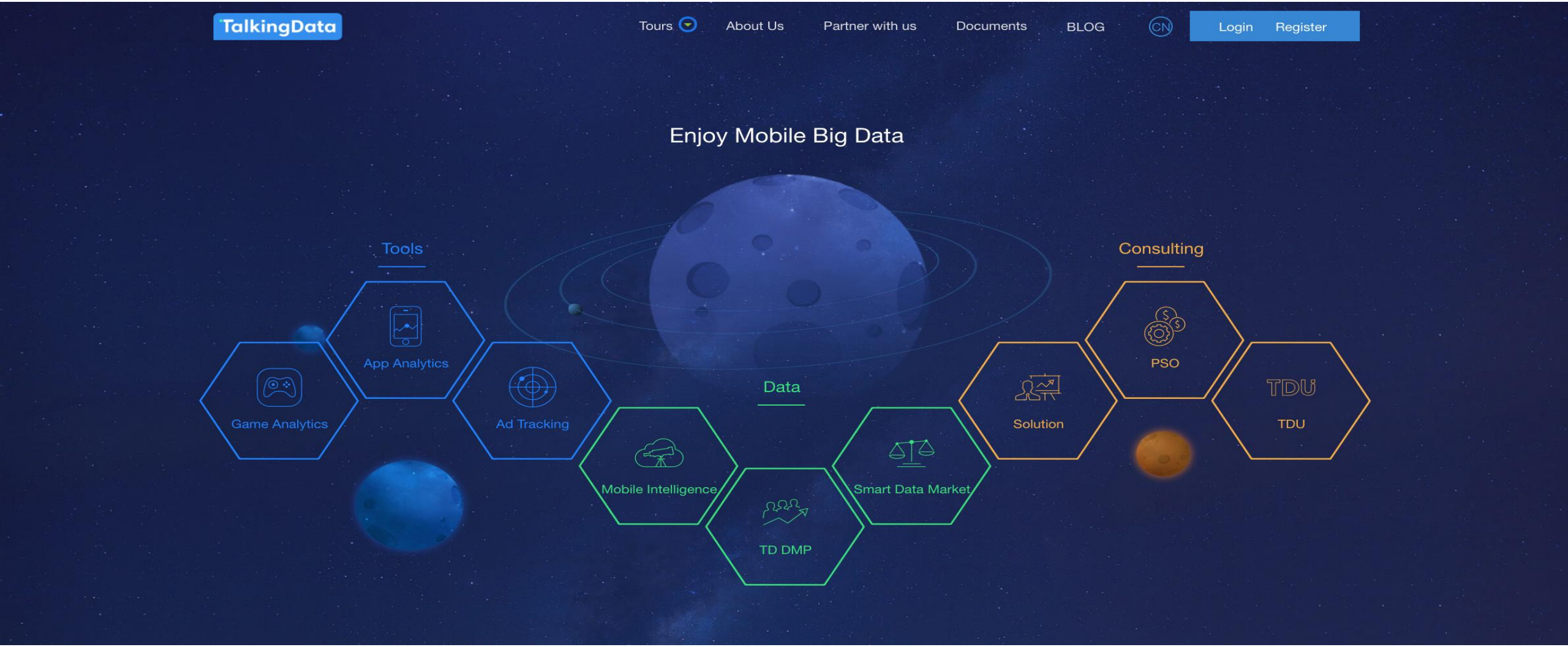
TalkingData AdTracking Fraud Detection Challenge

Can you detect fraudulent click traffic for mobile app ads?

[TalkingData]

China's largest independent big data service platform, covers over 70% of active mobile devices nationwide.



대회 설명

- 광고를 맡긴 회사가 원하는 것

광고를 보는 사람들이 광고를 클릭하고 그 앱을 다운로드 받는 행위까지 연결되는 것.

- 그러나, 광고를 클릭만 해도 광고를 맡긴 회사를 돈을 내야 한다.

- 광고를 클릭하기만 하고 다운 받지 않는 사람들을 사기라고 간주.

- 즉, 사기치는 사람들을 골라내자!

데이터 설명

train
ip
app
device
os
channel
click_time
is_attributed
attributed_time

test
click_id
ip
app
device
os
channel
click_time

비식별 처리로 인해
숫자로 주어짐
-> factor 변환

어려웠던 점(삽질기)

1. 데이터 로드

Mac Pro 8G RAM VS Windows 8G RAM

Windows 8G RAM → Microsoft Azure

어려웠던 점(삽질기)

1. 데이터 로드

Microsoft Azure (Rstudio Server)에서도 문제 발생

- Ds12 V2 (4 vcpu, 28G RAM)

데이터가 전부 읽어진다.
하지만 이후 진행이 불가..
(Cannot allocate 에러 발생)

- Ds13 V2 (8 vcpu, 56G RAM)

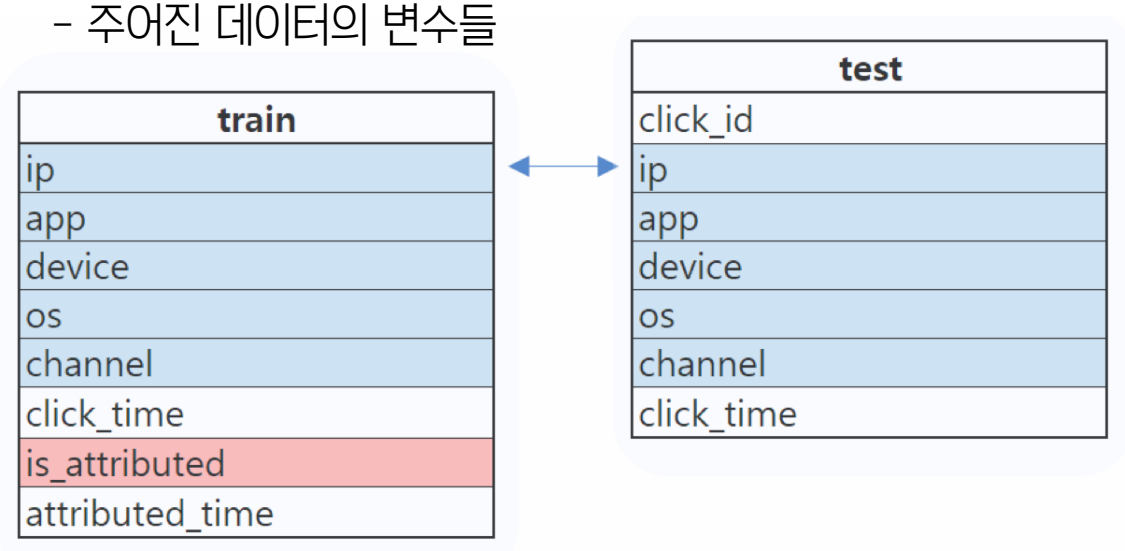
데이터가 전부 읽어지지 않음.
총 1억 8000만개의 데이터 중
1억 5000만개의 데이터만 로드됨.

데이터를 날짜별로 Subset하여 활용!!!

어려웠던 점(삽질기)

2. 데이터 이해

- 주어진 데이터의 변수들



train
ip
app
device
os
channel
click_time
is_attributed
attributed_time

test
click_id
ip
app
device
os
channel
click_time

- 그리고 처음 한 생각들

- ip, app, device, os, channe은 Factor다.
- is_attributed, attributed_time, click_time으로는 파생변수를 어떻게 만들지?
- 그냥 모델링 싸움인건가?

어려웠던 점(삽질기)

2. 데이터 이해

하지만 데이터를 조금 살펴보니,

```
> train[,c("day","hr")] %>% table
   hr
day       0       1       2       3       4       5       6       7       8       9      10      11      12      13      14      15      16      17      18      19      20      21      22      23
  6       0       0       0       0       0       0       0       0       0       0       0       0       0       0      48     435 2307777 1263348  738140  496354  409752  509572 1223530 2359612
  7 3604365 3308150 3095633 3220271 3645493 3227349 2917284 2924033 2801293 2961319 3300746 3134200 3000091 3235239 3161797 2864021 2314780 1263077  726684  493056  409255  518866 1172535 2333773
  8 3493769 3065649 3585843 3172056 3545132 3160269 2983655 3155262 2976057 3068314 3377086 3430977 3485357 3616634 3676695 3336168 2455567 1387383  794088  554053  447324  550518 1260525 2366694
  9 3318301 3082862 3068887 3351149 4032691 3671741 3570940 3186240 2804701 2986204 3304199 3347741 3363917 3457523 3443283 3026111     447       0       0       0       0       0       0       0
```

Train data의 날짜는 2017-11-06 ~ 2017-11-09. 총 4일.
Test data의 날짜는 2017-11-10이다.

결국 과거 4일(6,7,8,9)의 데이터를 가지고 미래(10일)를 예측하는 것이었다!!

그런데...

시간이 2017-11-06 14시부터 2017-11-09 15시까지만 있다. 왜지???

중국시간 기준 이기 때문이었다. 주어진 데이터의 시간은 UTC + 0이지만,
중국은 UTC + 8이다. 따라서 8시간을 더해줘야 중국시간이 되는 것이다.

어려웠던 점(삽질기)

2. 데이터 이해

```
> train_7$hr %>% table
.
      0       1       2       3       4       5       6       7       8       9      10      11      12      13      14      15      16      17      18      19      20      21      22      23
2307777 1263348  738140  496354  409752  509572 1223530 2359612 3604365 3308150 3095633 3220271 3645493 3227349 2917284 2924033 2801293 2961319 3300746 3134200 3000091 3235239 3161797 2864021
> train_8$hr %>% table
.
      0       1       2       3       4       5       6       7       8       9      10      11      12      13      14      15      16      17      18      19      20      21      22      23
2314780 1263077  726684  493056  409255  518866 1172535 2333773 3493769 3065649 3585843 3172056 3545132 3160269 2983655 3155262 2976057 3068314 3377086 3430977 3485357 3616634 3676695 3336168
> train_9$hr %>% table
.
      0       1       2       3       4       5       6       7       8       9      10      11      12      13      14      15      16      17      18      19      20      21      22      23
2455567 1387383  794088  554053  447324  550518 1260525 2366694 3318301 3082862 3068887 3351149 4032691 3671741 3570940 3186240 2804701 2986204 3304199 3347741 3363917 3457523 3443283 3026111
> test$hr %>% table
.
     12      13      14      17      18      19      21      22      23
3344125 2858427     381 2984808 3127993     413 3212566 3261257     499
```

참고할 점 >

Competition에 참가하기 전, 처음 올라왔던 Test data가 삭제되고 새로운 Test data로 변경되었는데,
Old Test data에는 시간이 0시부터 23시까지 전부있다..

이 Old Test data를 어떻게 활용할 수 있을지 고민해 볼 필요가 있을것.

EDA

- 타겟 변수 : is_attributed => **Unbalanced problem**

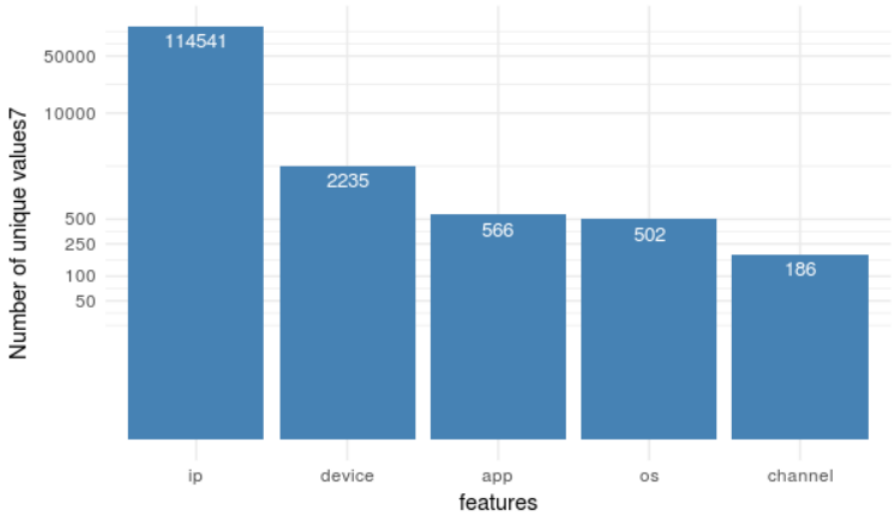
DAY \ Is_attributed	0	1
7일	0.997453	0.002547
8일	0.997572	0.002427
9일	0.997558	0.002441

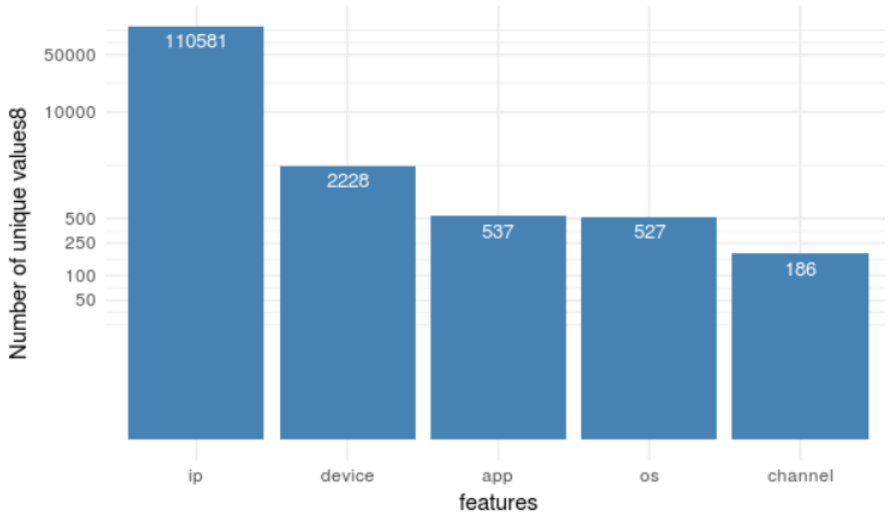
* 전체 변수에서 Missing Value는 없었음.

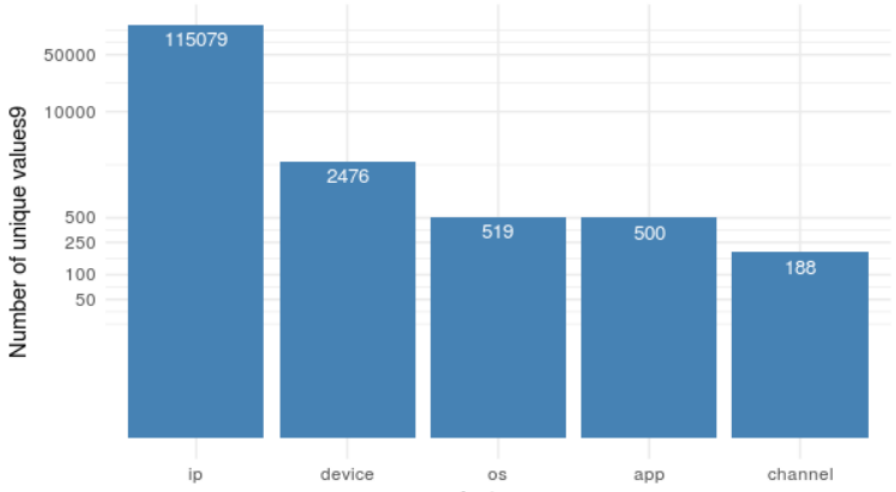
EDA

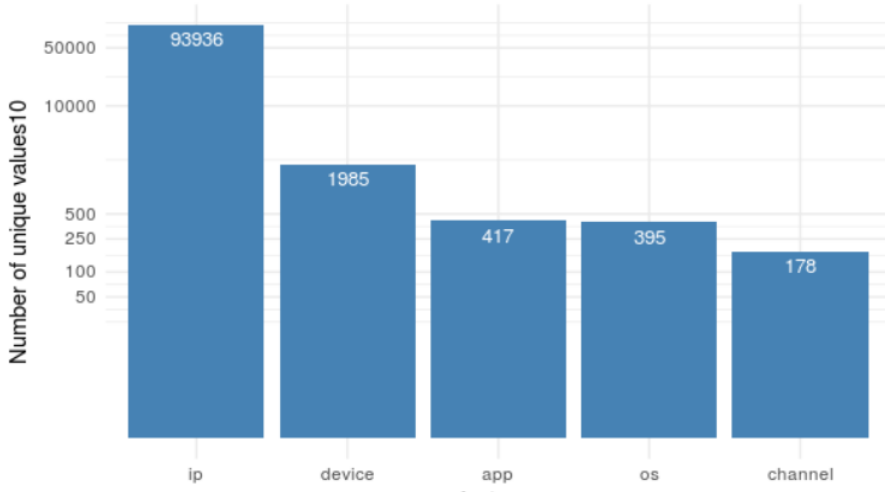
FACTOR의 LEVEL수

- 시간 변화를 Validation으로 사용하기 위하여 일별 분포를 살펴보자.





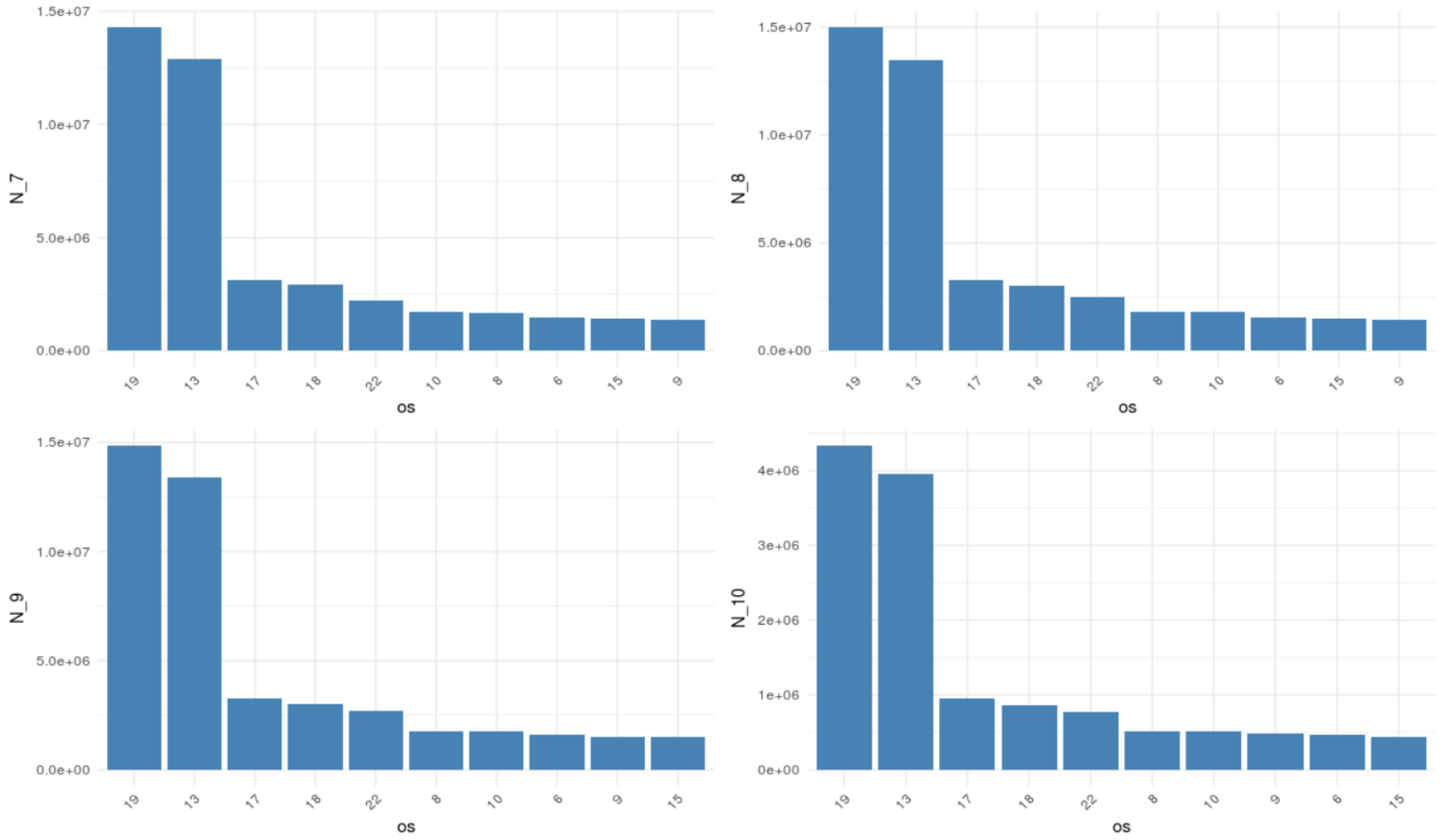




EDA

OS의 LEVEL별 빈도

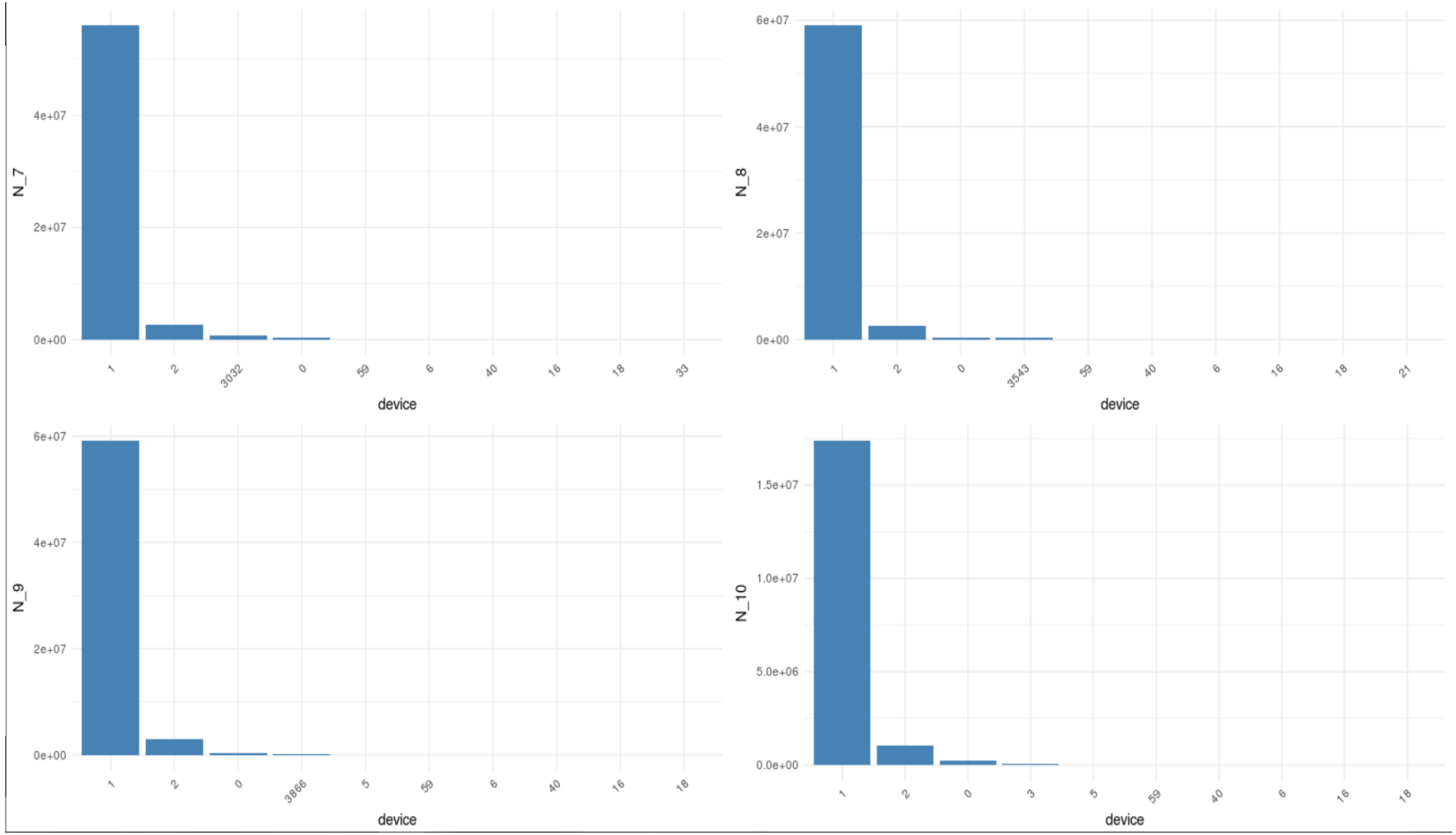
- 시간 변화를 Validation으로 사용하기 위하여 일별 분포를 살펴보자.



EDA

DEVICE의 LEVEL별 빈도

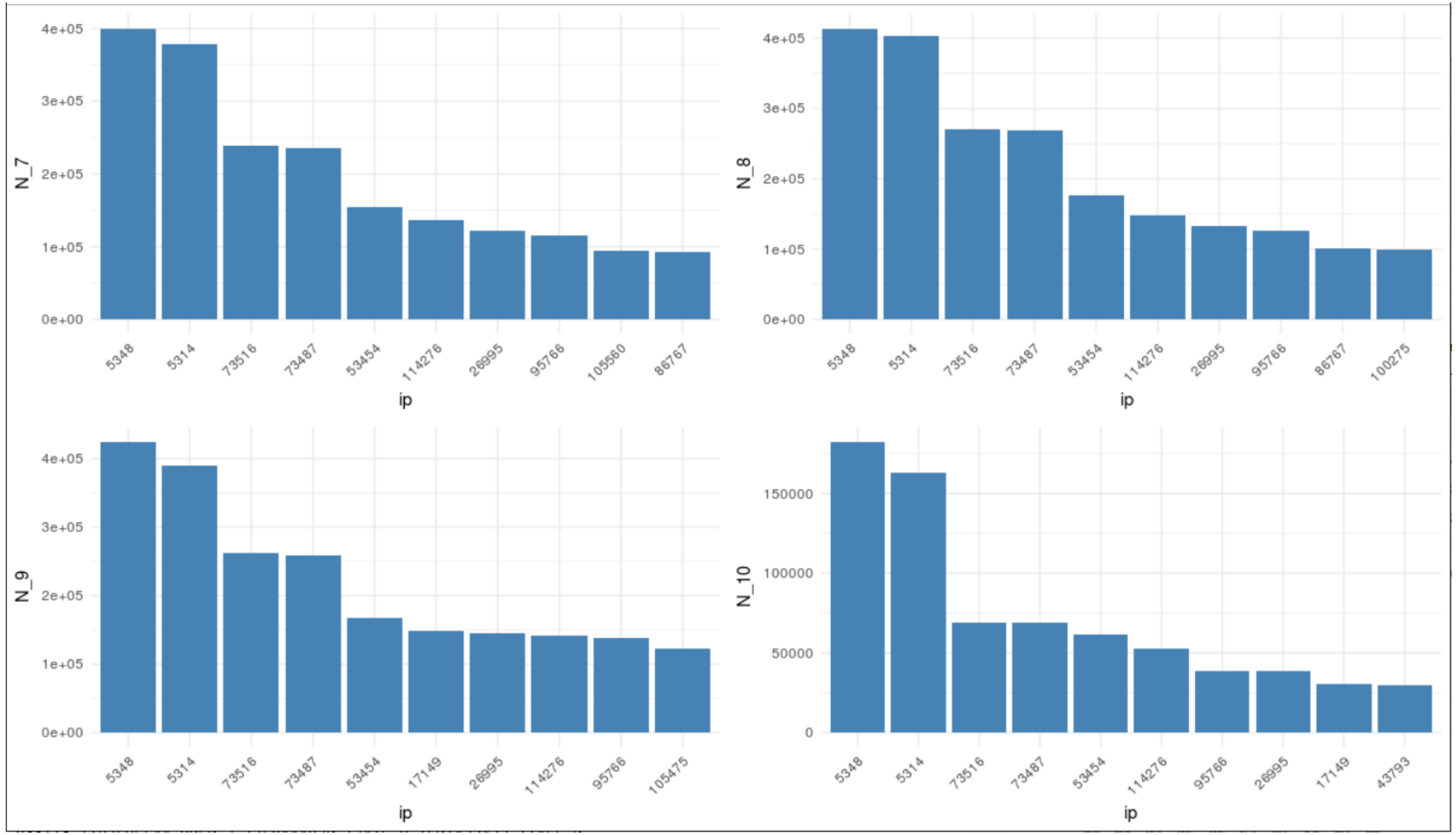
- 시간 변화를 Validation으로 사용하기 위하여 일별 분포를 살펴보자.



EDA

IP의 LEVEL별 빈도

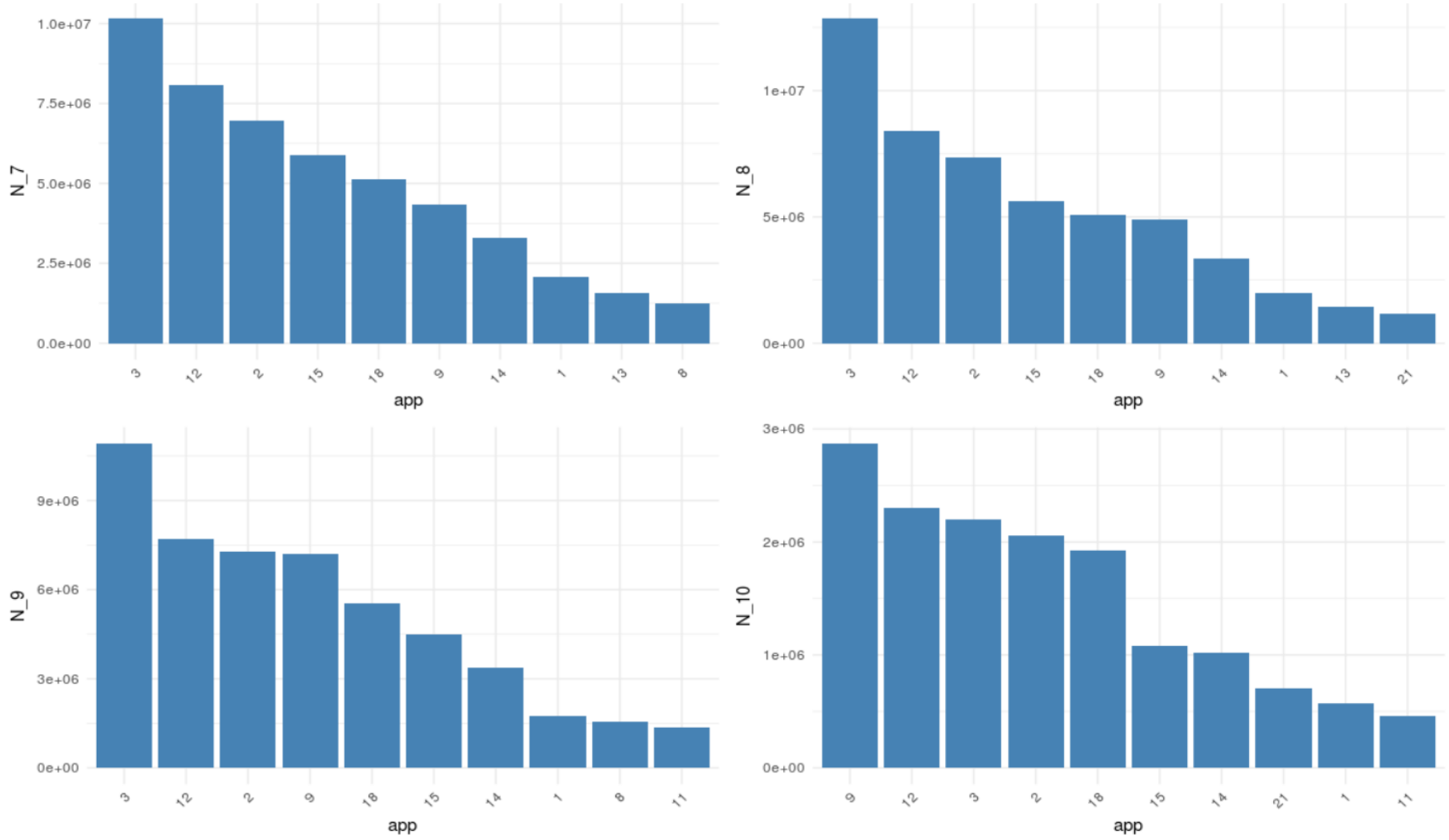
- 시간 변화를 Validation으로 사용하기 위하여 일별 분포를 살펴보자.



EDA

APP의 LEVEL별 빈도

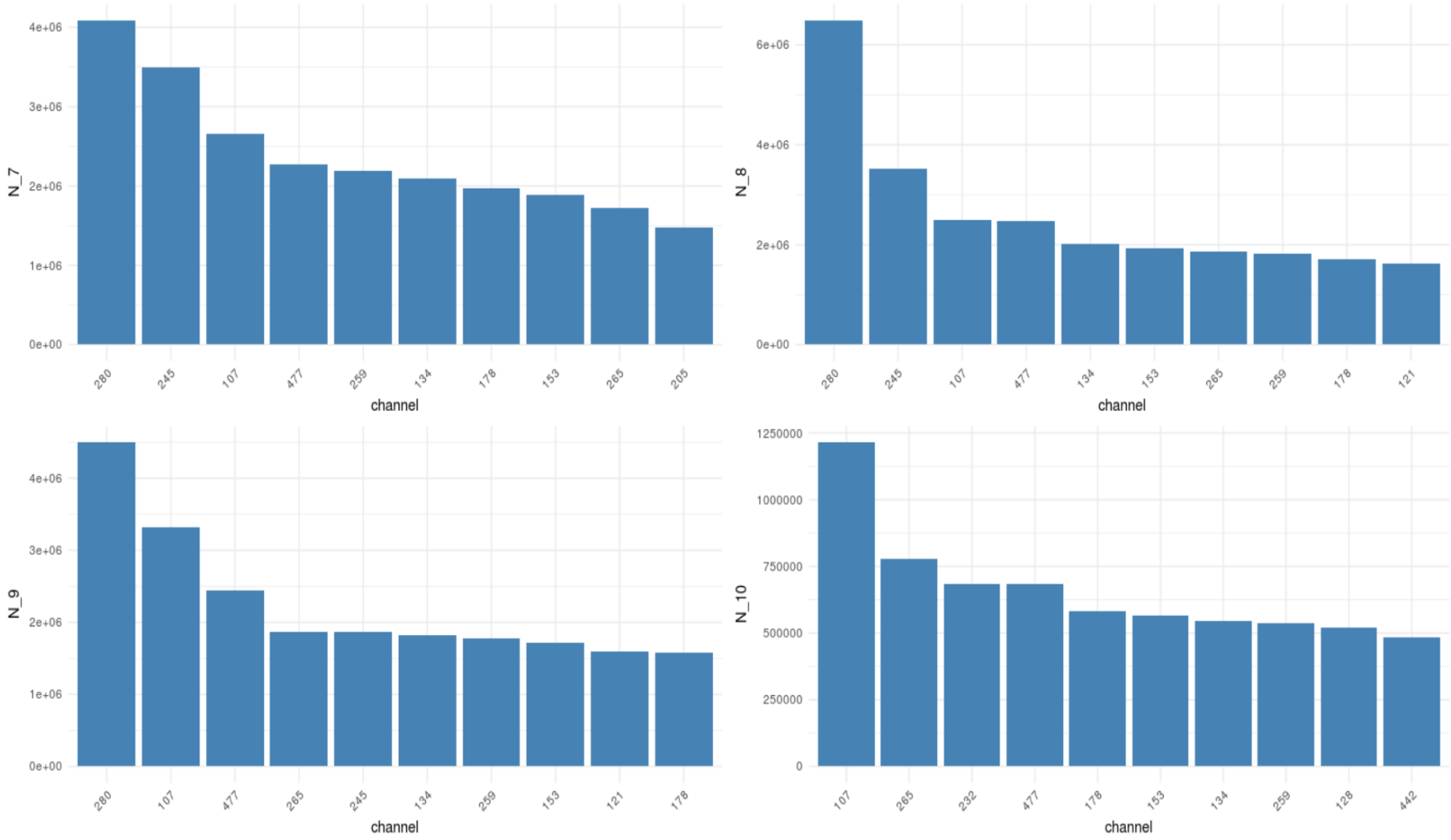
- 시간 변화를 Validation으로 사용하기 위하여 일별 분포를 살펴보자.



EDA

CHANNEL의 LEVEL별 빈도

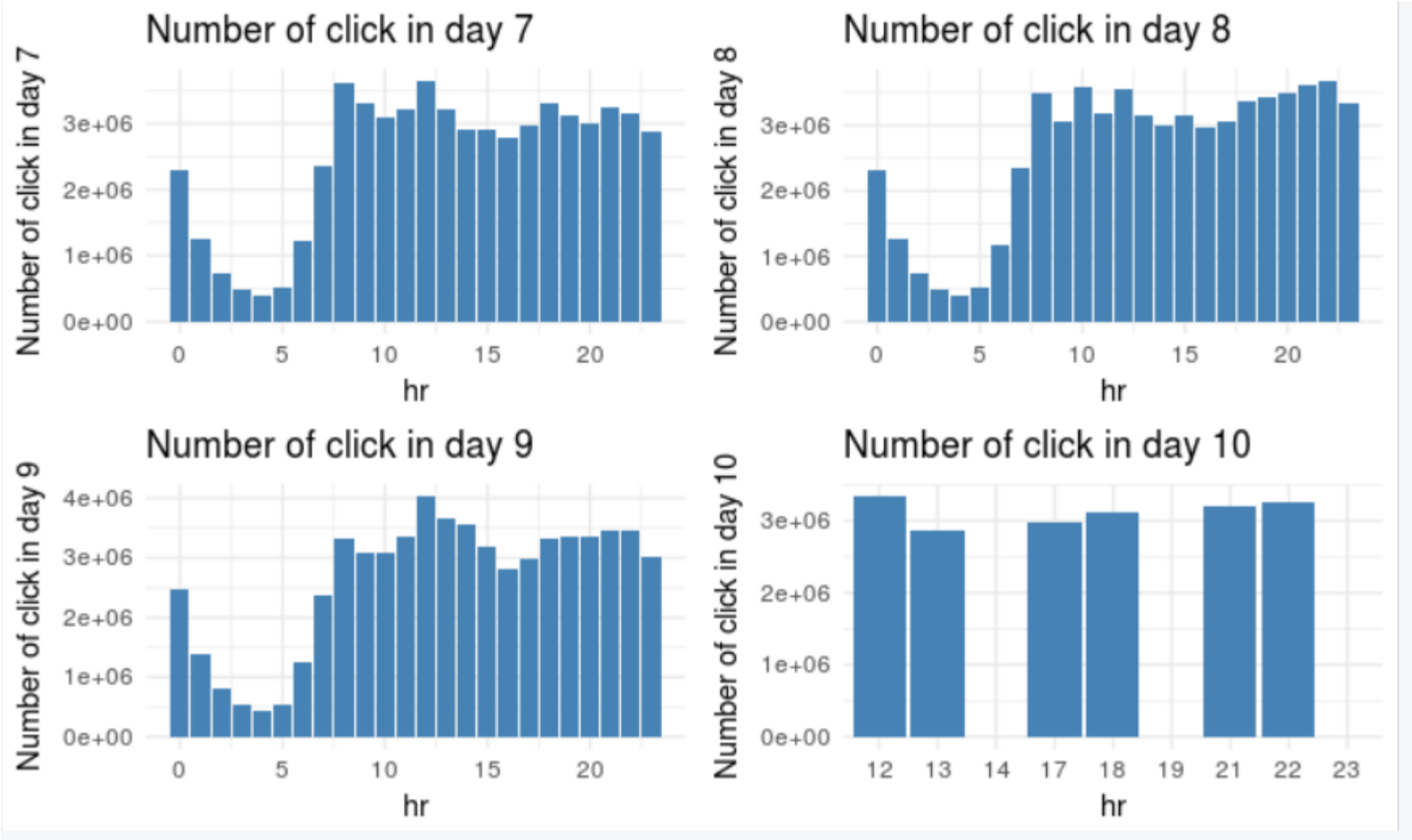
- 시간 변화를 Validation으로 사용하기 위하여 일별 분포를 살펴보자.



EDA

일별 클릭수

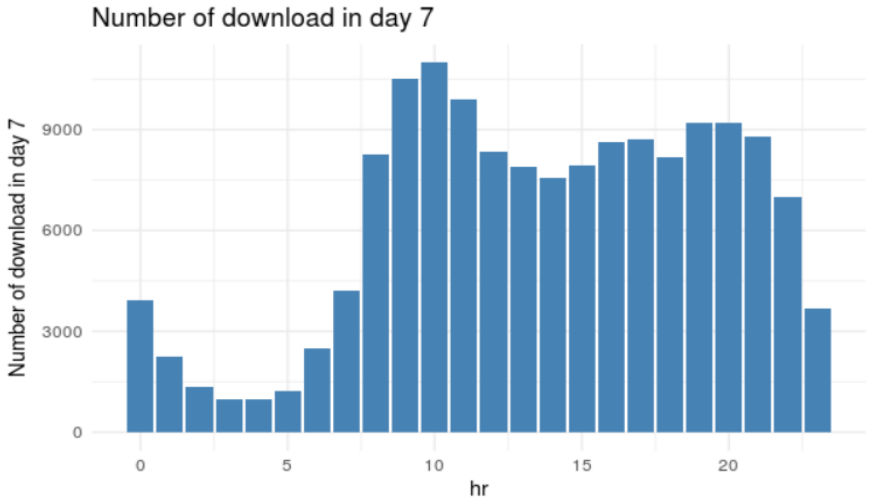
- 시간 변화를 Validation으로 사용하기 위하여 일별 분포를 살펴보자.

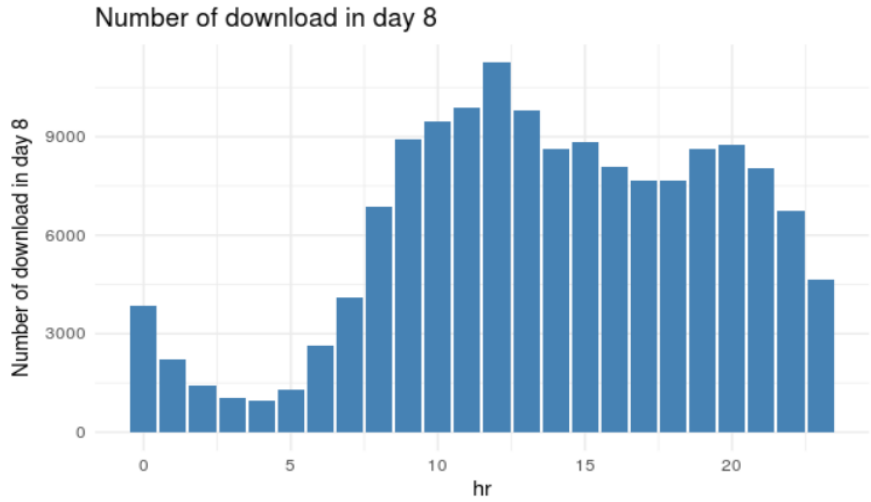


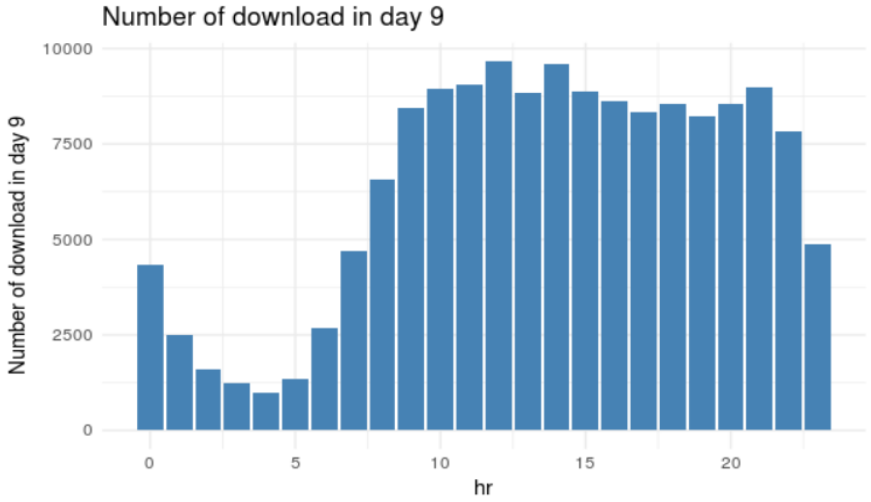
EDA

시간별 다운로드수

- 시간 변화를 Validation으로 사용하기 위하여 일별 분포를 살펴보자.







EDA 결론

1. 시간변화를 validation으로 활용가능하다는 것을 확인

2. 주의해야할 점 확인.

3. 파생변수의 아이디어 얻음.

추후 계획

1. 삽질 EDA 기반으로 파생변수 생성

2. 다운 샘플링 시도

3. 커널 참고하면서 모델링