07.31 ~ 08.06

데이터 다시 만들기 & threshold 다시 정하고, fraud 판별

A. full_train_no2_click&LDA_new.csv

- ~~500만개 읽고, total_rows 만큼 채움(대략 3200만개)~~
- `3698,0778` 개의 행

B. test

- `1879,0469` 개의 행

A+B : `5577,1247` 행

- 데이터프레임 합친 후, 뒤에 LDA부분만 따로 KMeans, 그룹 넘버를 `group` 피쳐로 추가

- threshold = 0.000057
- `13114` IP 유저 / `262421` 전체 유저 = 4.997%

```
13114 4.997313477198852%
```

- fraud에 해당하는 IP 리스트를 fr_ip, 0805_fraud_ip_list.pkl 에 저장해둠

시각화 1: 20개의 그룹마다, 1시간마다 각각의 fraud 비율 곡선

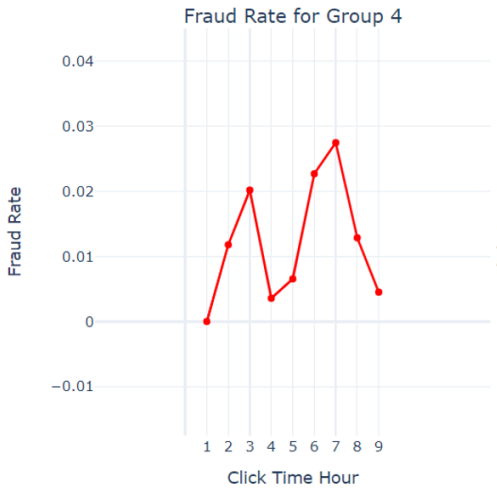
- 각각의 그룹 구성 클릭 개수

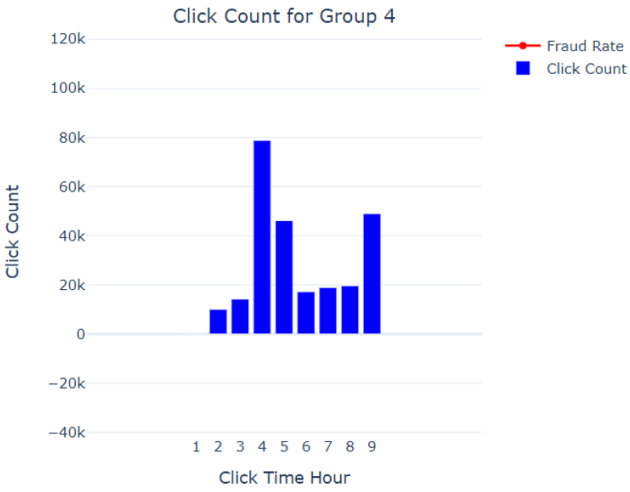
```
gr0: 3714870 gr1: 1664541 gr2: 569806 gr3: 3971843 gr4: 253537 gr5: 35
28857 gr6: 1607467 gr7: 2192586 gr8: 3231034 gr9: 3886468 gr10: 429156
6 gr11: 2205131 gr12: 2547228 gr13: 1246036 gr14: 3637732 gr15: 245906
5 gr16: 2062572 gr17: 4084628 gr18: 2897958 gr19: 5718322
```

I. 클릭 개수 그래프:

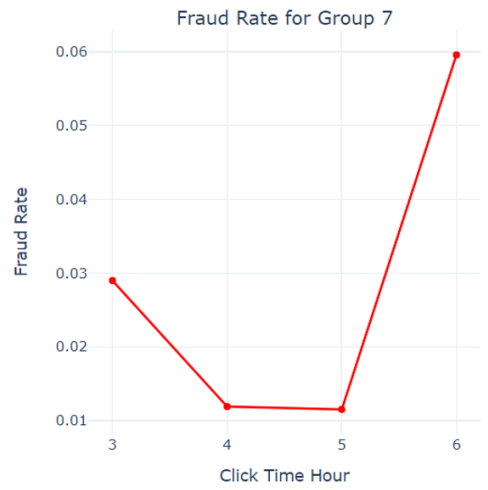
왼: 클릭 중 fraud 비율, 오: 클릭 개수

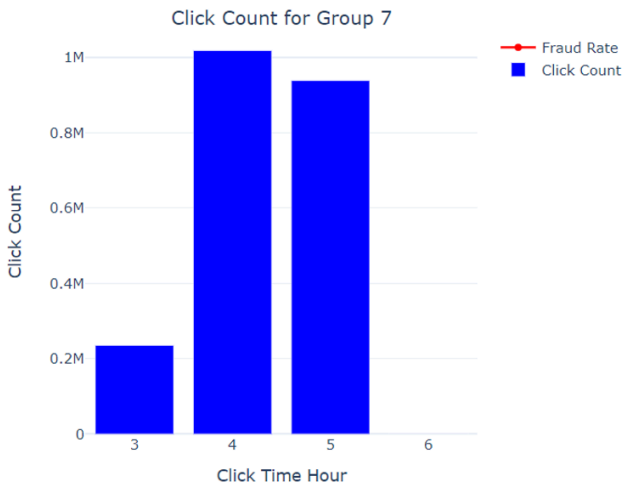
Group 4



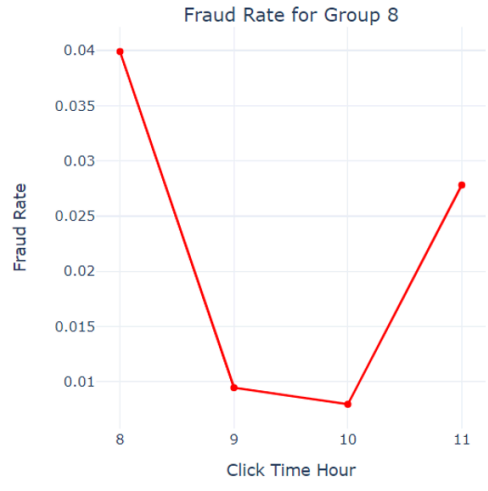


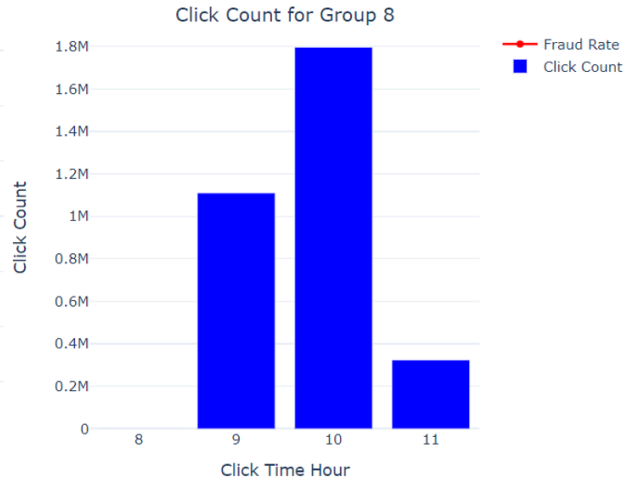
Group 7



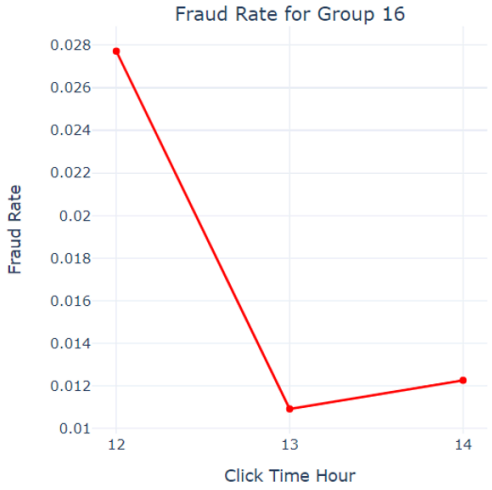


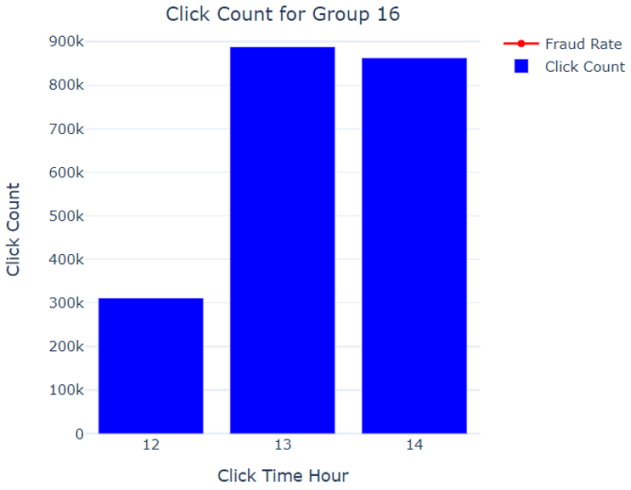
Group 8



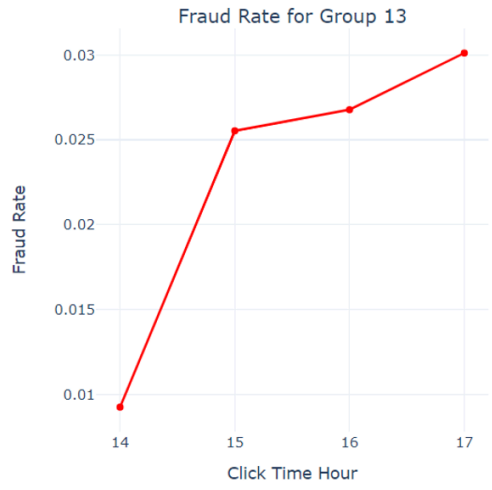


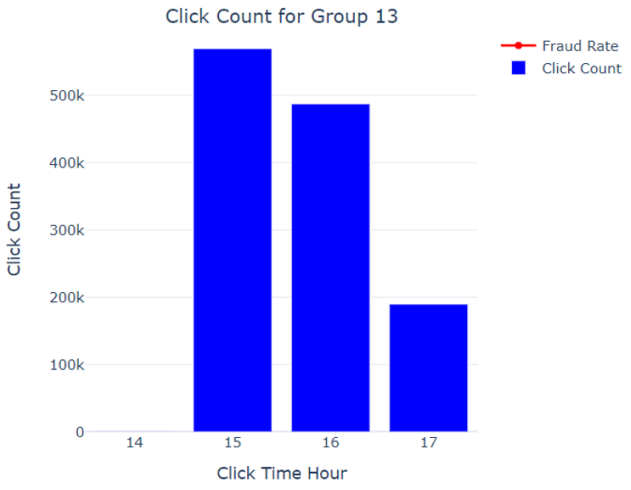
Group 16



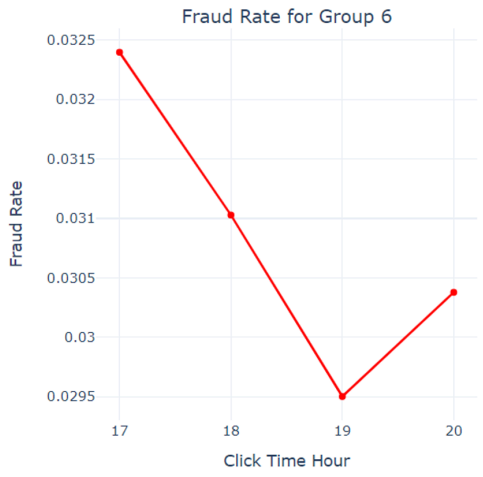


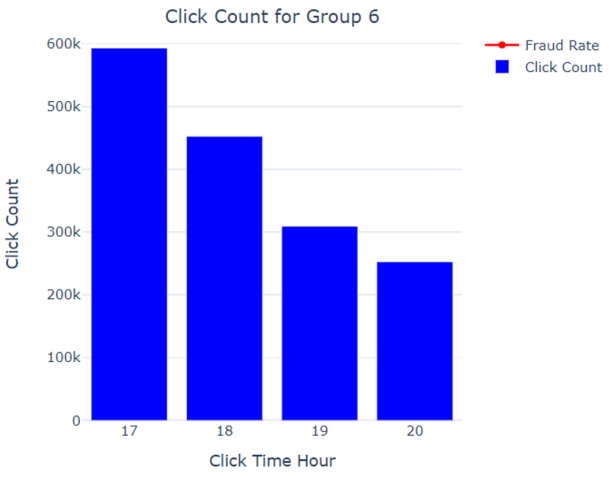
Group 13

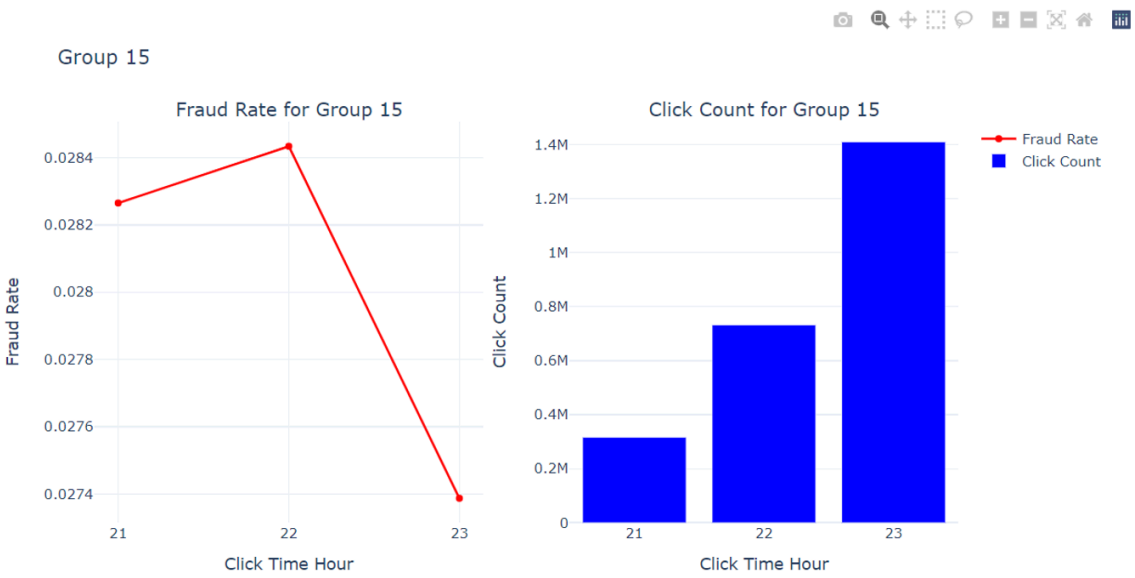




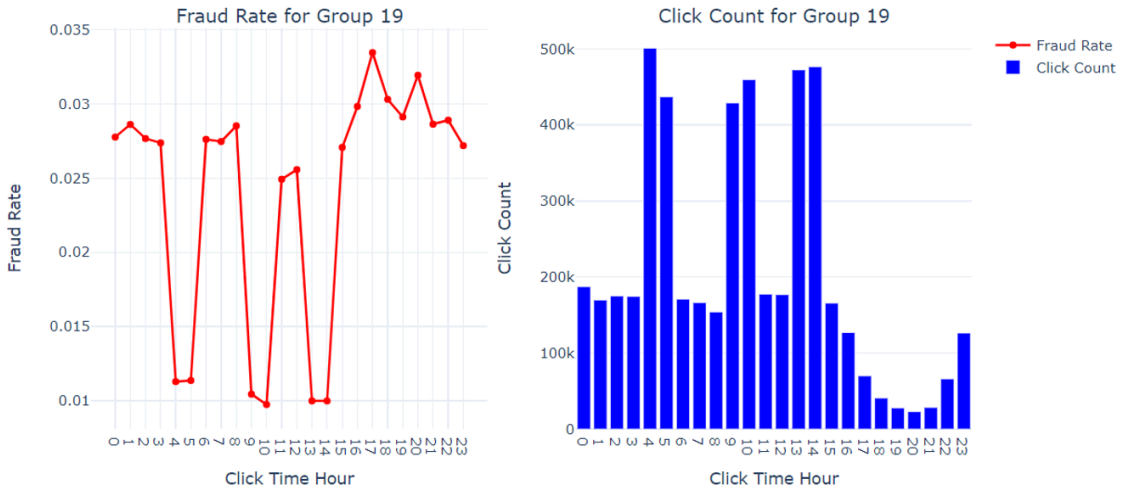
Group 6



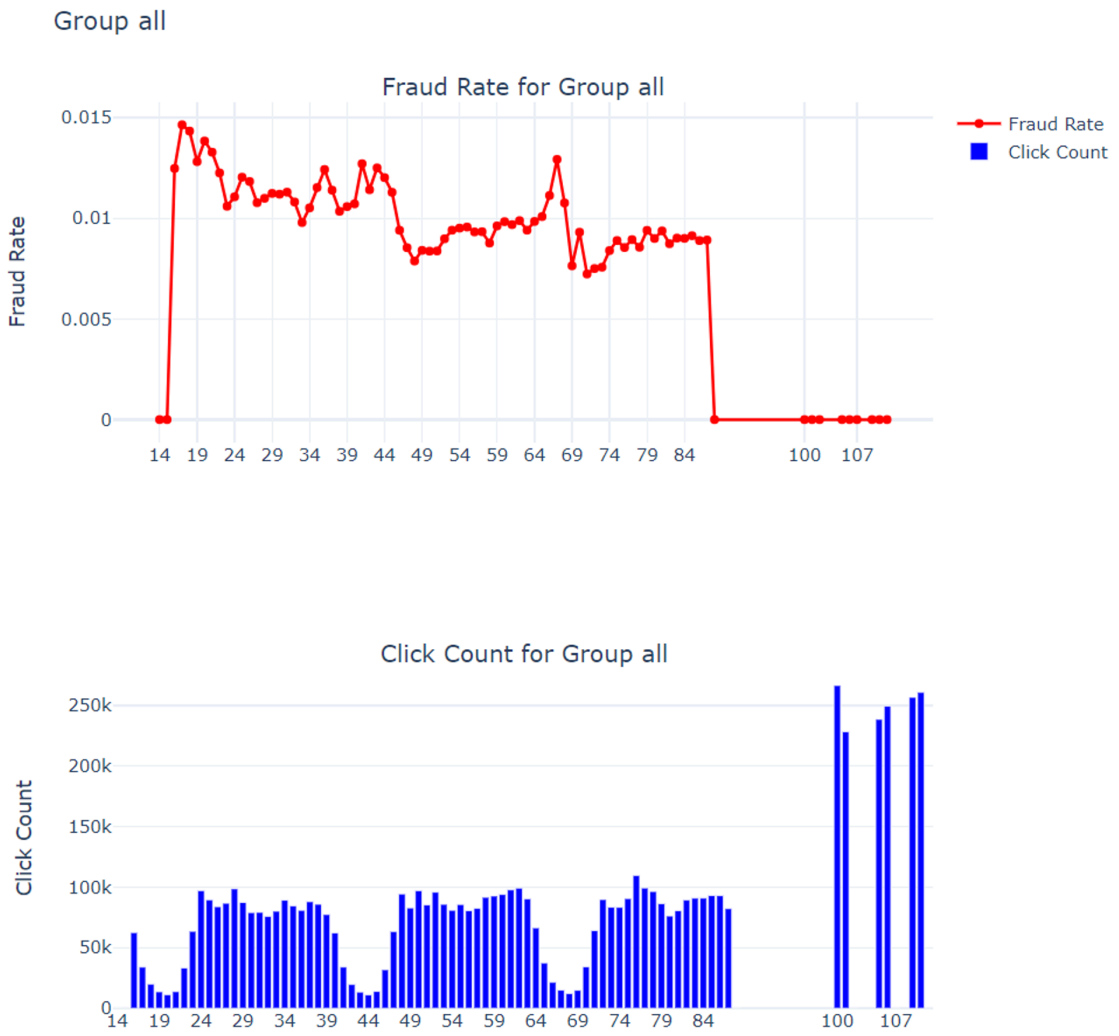




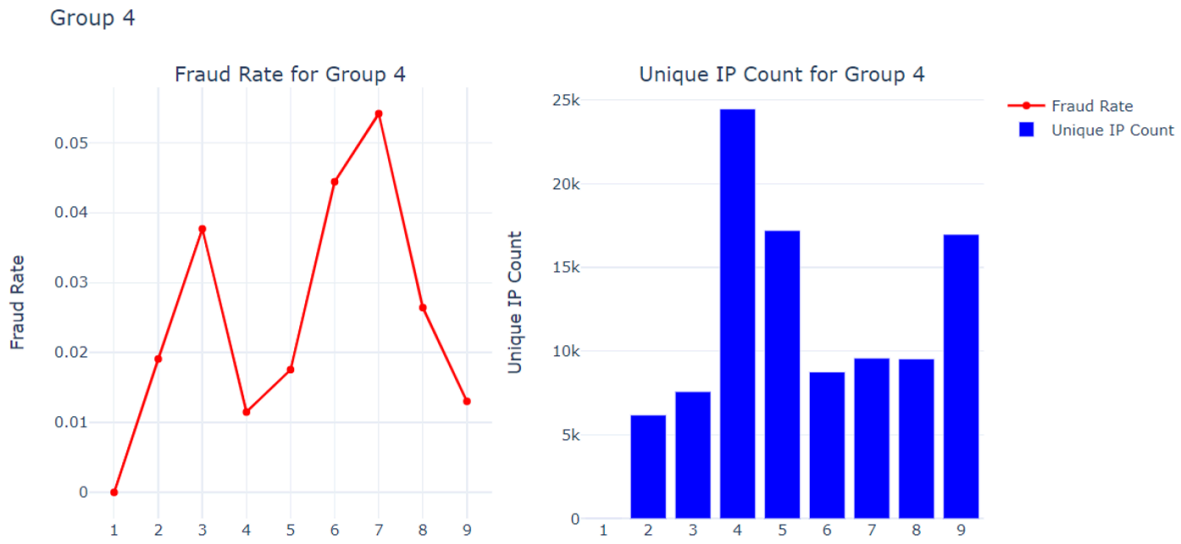
- 그룹 고려하지 않고 랜덤 5,000,000만 개



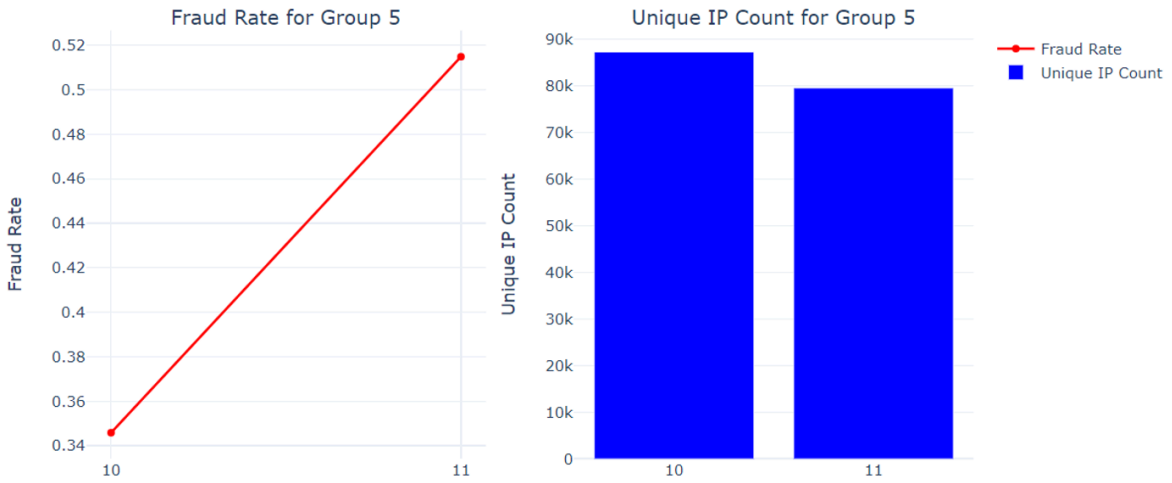
- 랜덤 650만개, 날짜(6~10일) 고려한 시간
  - (day - 6) * 24 + hour


Group all
Fraud Rate for Group all
Fraud Rate
0.015
0.01
0.005
0
14 19 24 29 34 39 44 49 54 59 64 69 74 79 84 100 107
Fraud Rate
Click Count
Click Count for Group all
Click Count
250k
200k
150k
100k
50k
0
14 19 24 29 34 39 44 49 54 59 64 69 74 79 84 100 107

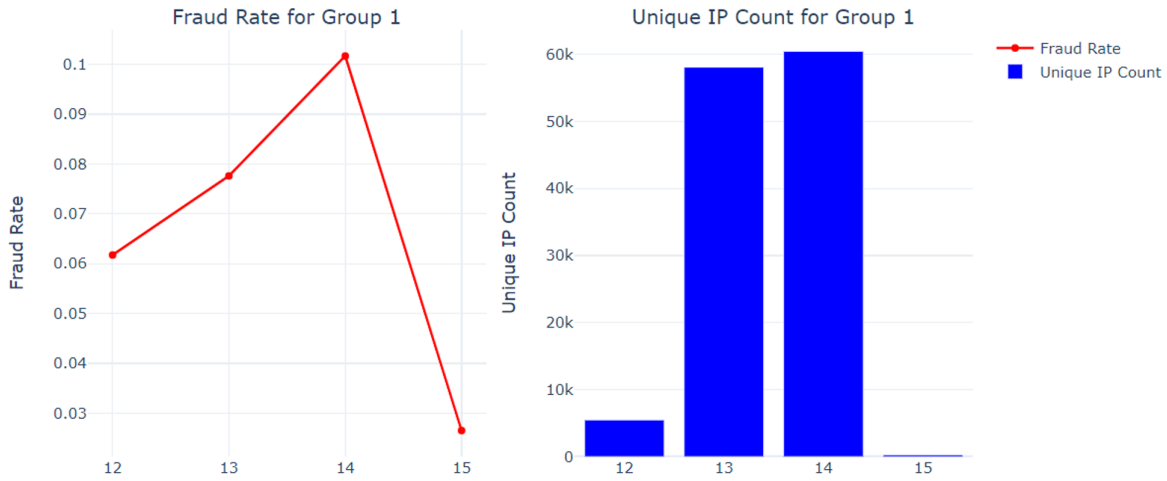
II. IP 유저 수 카운트


Group 4
Fraud Rate for Group 4
Fraud Rate
0.05
0.04
0.03
0.02
0.01
0
1 2 3 4 5 6 7 8 9
Unique IP Count for Group 4
Unique IP Count
25k
20k
15k
10k
5k
0
1 2 3 4 5 6 7 8 9
Fraud Rate
Unique IP Count

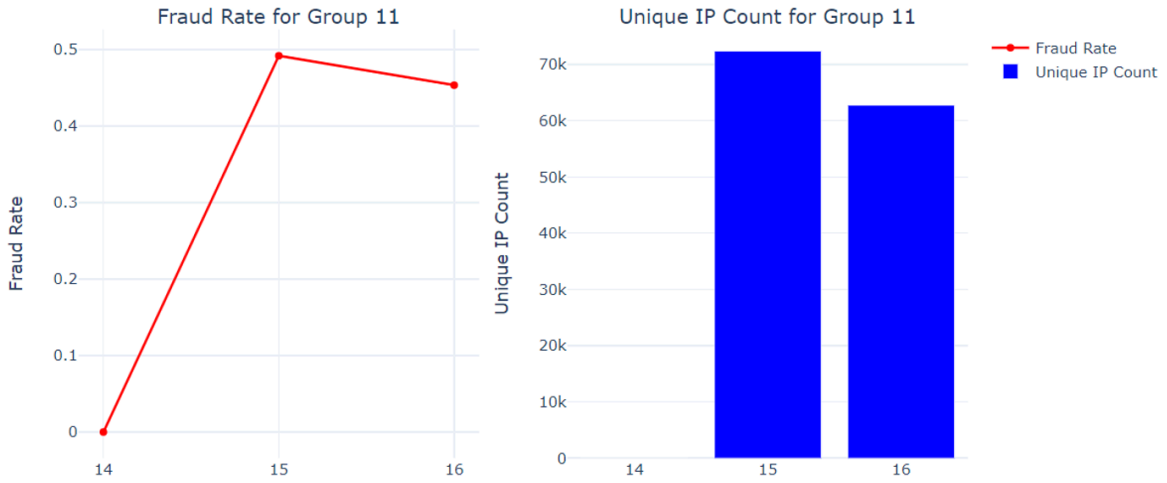
Group 5



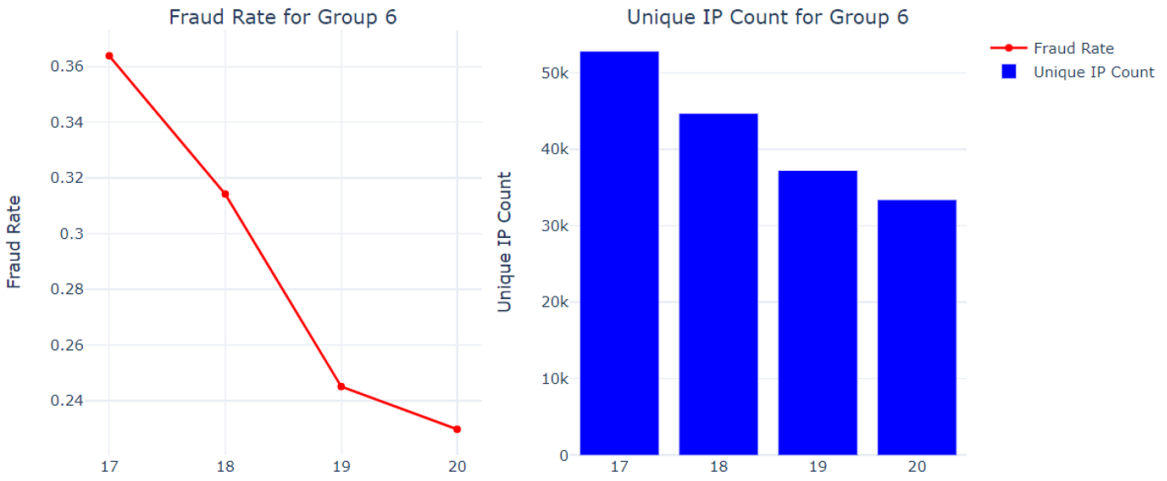
Group 1



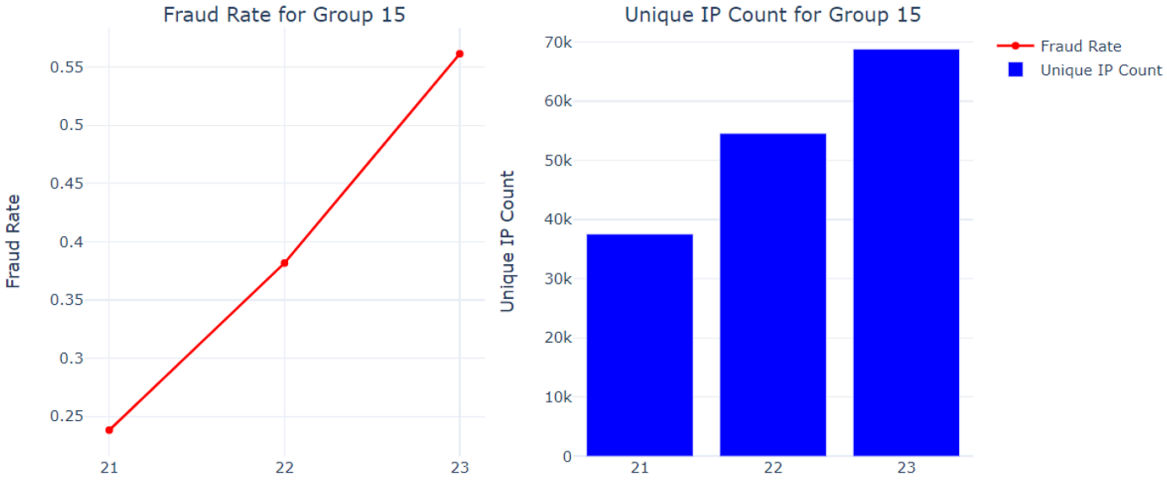
Group 11



Group 6



Group 15



- 650만개 샘플

글을 작성하거나 AI를 사용하려면 '스페이스' 키를, 명령어를 사용하려면 ' / ' 키를 누르세요.

시각화 2: 검사 룰을 적용했을 때 성과

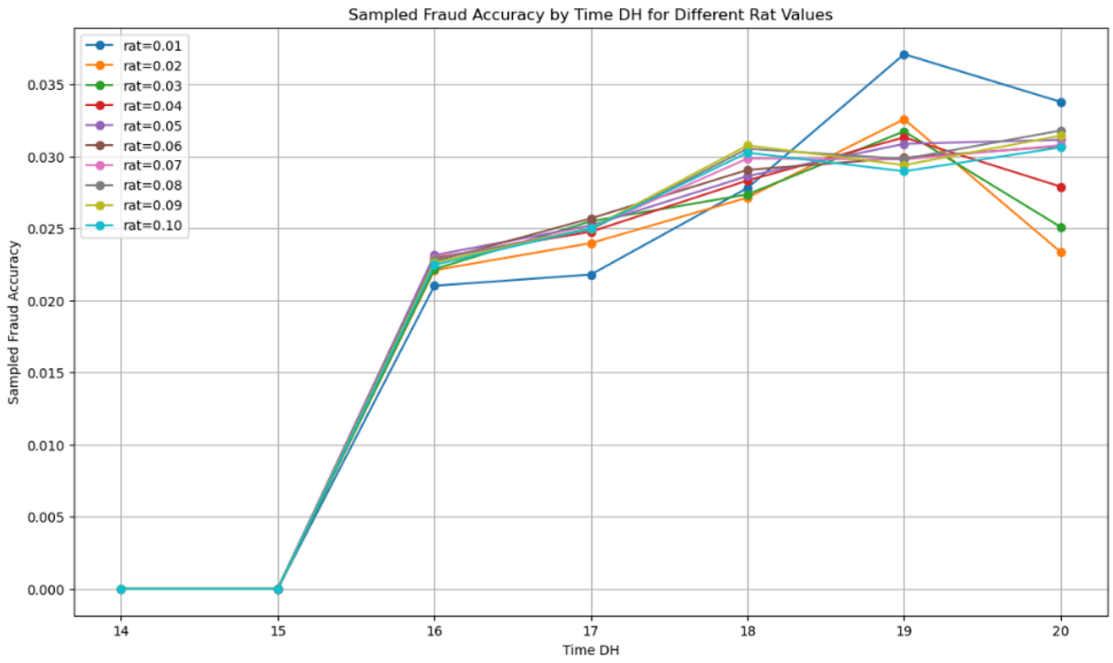
- 룰
  - 무지성으로 시간 단위로 (1 ~ 2% 검사를 하겠다) 유니폼하게, 기튼 지수 없이
  - 코로나 논문과 차이점: 앞으로 몇 명이 나올지 모름
    - 지난 한 시간에 클릭 수를 바탕으로 → 아직 미적용(클릭 수를 알고 있다고 가정하고)
  - 끝까지 블락(or 하루 블락) → 아직 미적용

- 확인한 값
  - 걸러지는 클릭 / 유저 수, 정확도
  - 놓친 fraud, 잡아낸 fraud 클릭/유저 수

```
Click Time Hour: 0
accuracy: 0.027262682551883168
Sampled Fraud Count: 1135  Ratio: 1.93%
Non-Sampled Fraud Count: 57767  Ratio: 98.07%
Sampled Unique IP Count: 21263
Sampled Fraud Unique IP Count: 862  Ratio: 0.04053990499929455
Non-Sampled Unique IP Count: 77889
Non-Sampled Fraud Unique IP Count: 5113  Ratio : 0.06564469950827459
--------------------------------------------------
Click Time Hour: 1
accuracy: 0.029228937971167834
Sampled Fraud Count: 1105  Ratio: 1.99%
Non-Sampled Fraud Count: 54454  Ratio: 98.01%
Sampled Unique IP Count: 20365
Sampled Fraud Unique IP Count: 859  Ratio: 0.042180211146575004
Non-Sampled Unique IP Count: 78895
Non-Sampled Fraud Unique IP Count: 5160  Ratio : 0.06540338424488244
--------------------------------------------------
Click Time Hour: 2
accuracy: 0.027894574915393294
Sampled Fraud Count: 1088  Ratio: 1.98%
Non-Sampled Fraud Count: 53952  Ratio: 98.02%
Sampled Unique IP Count: 20869
Sampled Fraud Unique IP Count: 845  Ratio: 0.04049067995591547
Non-Sampled Unique IP Count: 80382
Non-Sampled Fraud Unique IP Count: 5183  Ratio : 0.06447960986290463
--------------------------------------------------
Click Time Hour: 3
accuracy: 0.02737404344923219
Sampled Fraud Count: 1066  Ratio: 1.97%
Non-Sampled Fraud Count: 52948  Ratio: 98.03%
Sampled Unique IP Count: 20987
Sampled Fraud Unique IP Count: 835  Ratio: 0.03978653452137037
Non-Sampled Unique IP Count: 80863
Non-Sampled Fraud Unique IP Count: 5286  Ratio : 0.0653698230340205
--------------------------------------------------
Click Time Hour: 4
accuracy: 0.011201073585327667
Sampled Fraud Count: 1252  Ratio: 1.98%
Non-Sampled Fraud Count: 61922  Ratio: 98.02%
Sampled Unique IP Count: 36284
Sampled Fraud Unique IP Count: 947  Ratio: 0.02609965825157094
Non-Sampled Unique IP Count: 106121
Non-Sampled Fraud Unique IP Count: 5580  Ratio : 0.05258148717030559
--------------------------------------------------
Click Time Hour: 5
accuracy: 0.011202842297660855
Sampled Fraud Count: 1091  Ratio: 1.95%
Non-Sampled Fraud Count: 54916  Ratio: 98.05%
Sampled Unique IP Count: 34136
Sampled Fraud Unique IP Count: 867  Ratio: 0.025398406374501994
Non-Sampled Unique IP Count: 103521
```

```
Non-Sampled Fraud Unique IP Count: 5463  Ratio : 0.052771901353348596

Click Time Hour: 6
accuracy: 0.029074518280306823
Sampled Fraud Count: 1103  Ratio: 2.08%
Non-Sampled Fraud Count: 51926  Ratio: 97.92%
Sampled Unique IP Count: 20609
Sampled Fraud Unique IP Count: 875  Ratio: 0.042457178902421275
Non-Sampled Unique IP Count: 81054
Non-Sampled Fraud Unique IP Count: 5348  Ratio : 0.06598070422187677
--------------------------------------------------
Click Time Hour: 7
accuracy: 0.028610721605004854
Sampled Fraud Count: 1061  Ratio: 2.05%
Non-Sampled Fraud Count: 50718  Ratio: 97.95%
Sampled Unique IP Count: 20119
Sampled Fraud Unique IP Count: 842  Ratio: 0.04185098662955415
Non-Sampled Unique IP Count: 80801
Non-Sampled Fraud Unique IP Count: 5334  Ratio : 0.06601403447977129
--------------------------------------------------
Click Time Hour: 8
accuracy: 0.028459188127455258
Sampled Fraud Count: 978  Ratio: 2.07%
Non-Sampled Fraud Count: 46337  Ratio: 97.93%
Sampled Unique IP Count: 19004
Sampled Fraud Unique IP Count: 780  Ratio: 0.041043990738791836
Non-Sampled Unique IP Count: 79082
Non-Sampled Fraud Unique IP Count: 5156  Ratio : 0.06519814875698642
--------------------------------------------------
Click Time Hour: 9
accuracy: 0.01058007645227999
Sampled Fraud Count: 1013  Ratio: 2.09%
Non-Sampled Fraud Count: 47408  Ratio: 97.91%
Sampled Unique IP Count: 32833
Sampled Fraud Unique IP Count: 793  Ratio: 0.024152529467304236
Non-Sampled Unique IP Count: 100694
Non-Sampled Fraud Unique IP Count: 5146  Ratio : 0.05110532901662462
--------------------------------------------------
Click Time Hour: 10
accuracy: 0.010079622184926428
Sampled Fraud Count: 1033  Ratio: 2.07%
Non-Sampled Fraud Count: 48911  Ratio: 97.93%
Sampled Unique IP Count: 33498
Sampled Fraud Unique IP Count: 825  Ratio: 0.024628336020060898
Non-Sampled Unique IP Count: 103290
Non-Sampled Fraud Unique IP Count: 5364  Ratio : 0.051931455126343305
--------------------------------------------------
```



검사 비율을 조절해줄 때, 시간에 따른 정확도

```
rat 0.01: ['nan', '0.0000', '0.0210', '0.0218', '0.0278', '0.0371', '0.0338']
rat 0.02: ['nan', '0.0000', '0.0221', '0.0240', '0.0271', '0.0326', '0.0233']
rat 0.03: ['nan', '0.0000', '0.0222', '0.0255', '0.0274', '0.0317', '0.0251']
rat 0.04: ['nan', '0.0000', '0.0230', '0.0248', '0.0283', '0.0313', '0.0279']
rat 0.05: ['nan', '0.0000', '0.0231', '0.0252', '0.0286', '0.0309', '0.0311']
rat 0.06: ['nan', '0.0000', '0.0227', '0.0257', '0.0290', '0.0299', '0.0307']
rat 0.07: ['nan', '0.0000', '0.0229', '0.0250', '0.0298', '0.0298', '0.0308']
rat 0.08: ['0.0000', '0.0000', '0.0229', '0.0249', '0.0305', '0.0298', '0.0318']
rat 0.09: ['0.0000', '0.0000', '0.0226', '0.0250', '0.0307', '0.0294', '0.0314']
rat 0.10: ['0.0000', '0.0000', '0.0225', '0.0250', '0.0302', '0.0290', '0.0306']
rat 0.11: ['0.0000', '0.0000', '0.0228', '0.0251', '0.0298', '0.0283', '0.0290']
rat 0.12: ['0.0000', '0.0000', '0.0230', '0.0251', '0.0300', '0.0278', '0.0298']
rat 0.13: ['0.0000', '0.0000', '0.0233', '0.0251', '0.0298', '0.0282', '0.0303']
rat 0.14: ['0.0000', '0.0000', '0.0235', '0.0254', '0.0297', '0.0279', '0.0298']
rat 0.15: ['0.0000', '0.0000', '0.0233', '0.0257', '0.0294', '0.0276', '0.0299']
```

```
--------------------------------------------------
Click Time Hour: 11
accuracy: 0.02569209823004387
Sampled Fraud Count: 1019  Ratio: 2.04%
Non-Sampled Fraud Count: 48929  Ratio: 97.96%
Sampled Unique IP Count: 19149
Sampled Fraud Unique IP Count: 790  Ratio: 0.04125541803749543
Non-Sampled Unique IP Count: 82136
Non-Sampled Fraud Unique IP Count: 5426  Ratio : 0.06606116684523229
--------------------------------------------------
Click Time Hour: 12
accuracy: 0.025517976843388177
Sampled Fraud Count: 1005  Ratio: 2.00%
Non-Sampled Fraud Count: 49182  Ratio: 98.00%
Sampled Unique IP Count: 18695
Sampled Fraud Unique IP Count: 777  Ratio: 0.04156191495052153
Non-Sampled Unique IP Count: 81897
Non-Sampled Fraud Unique IP Count: 5454  Ratio : 0.06659584600168504
```