데이터 시각화 실전 문제 풀이

In [15]:
```
library(dplyr)
library(ggplot2)
```

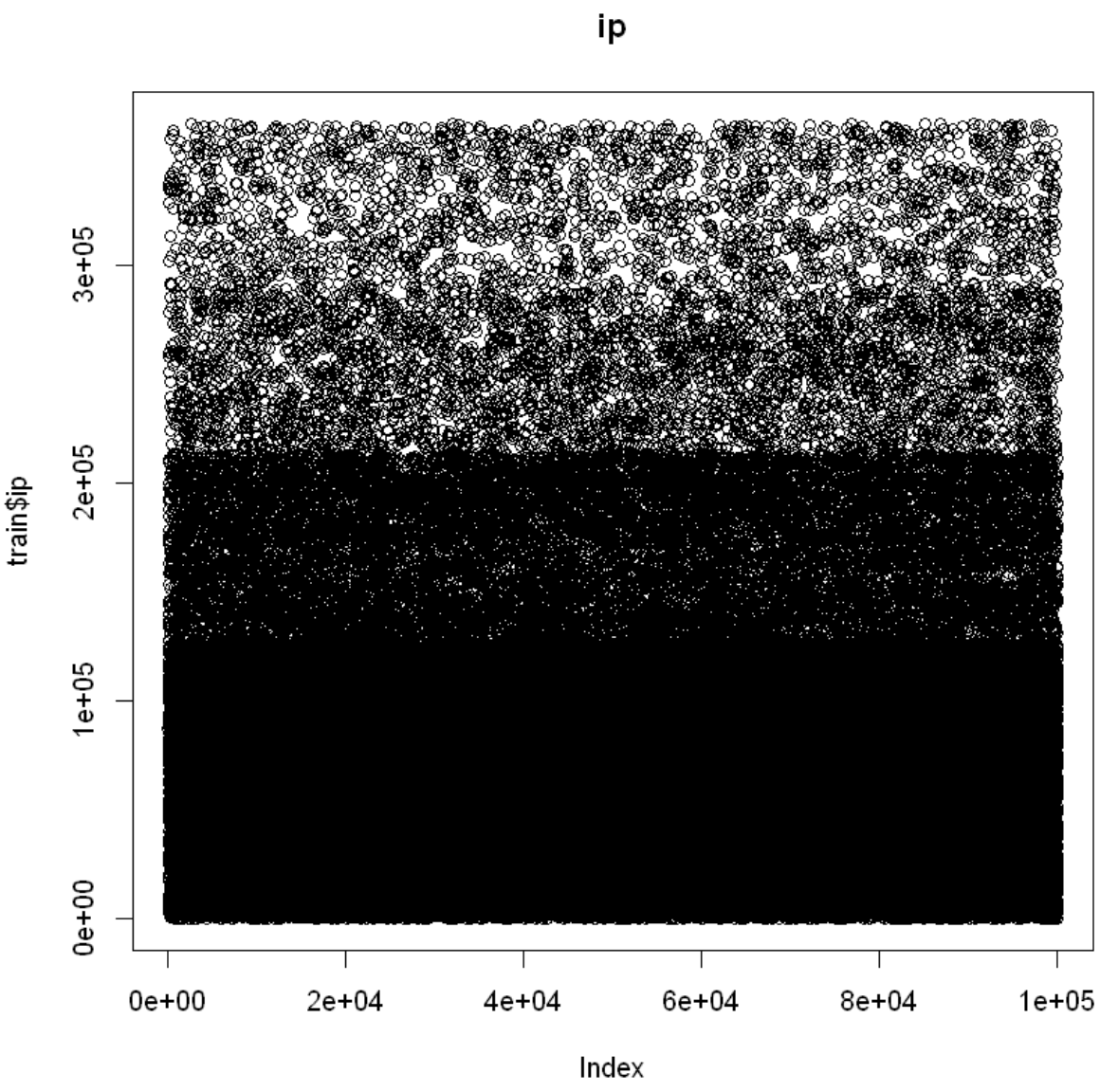
In [14]:
```
train = read.csv("C:/data/ad_tracking/train_sample.csv")
names(train)
table(train$is_attributed);
```

'ip' 'app' 'device' 'os' 'channel' 'click_time' 'attributed_time' 'is_attributed'

```
    0     1 
99773   227 
```

실습 과제1-1

In [5]:
```
plot(train$ip, main="ip")
```

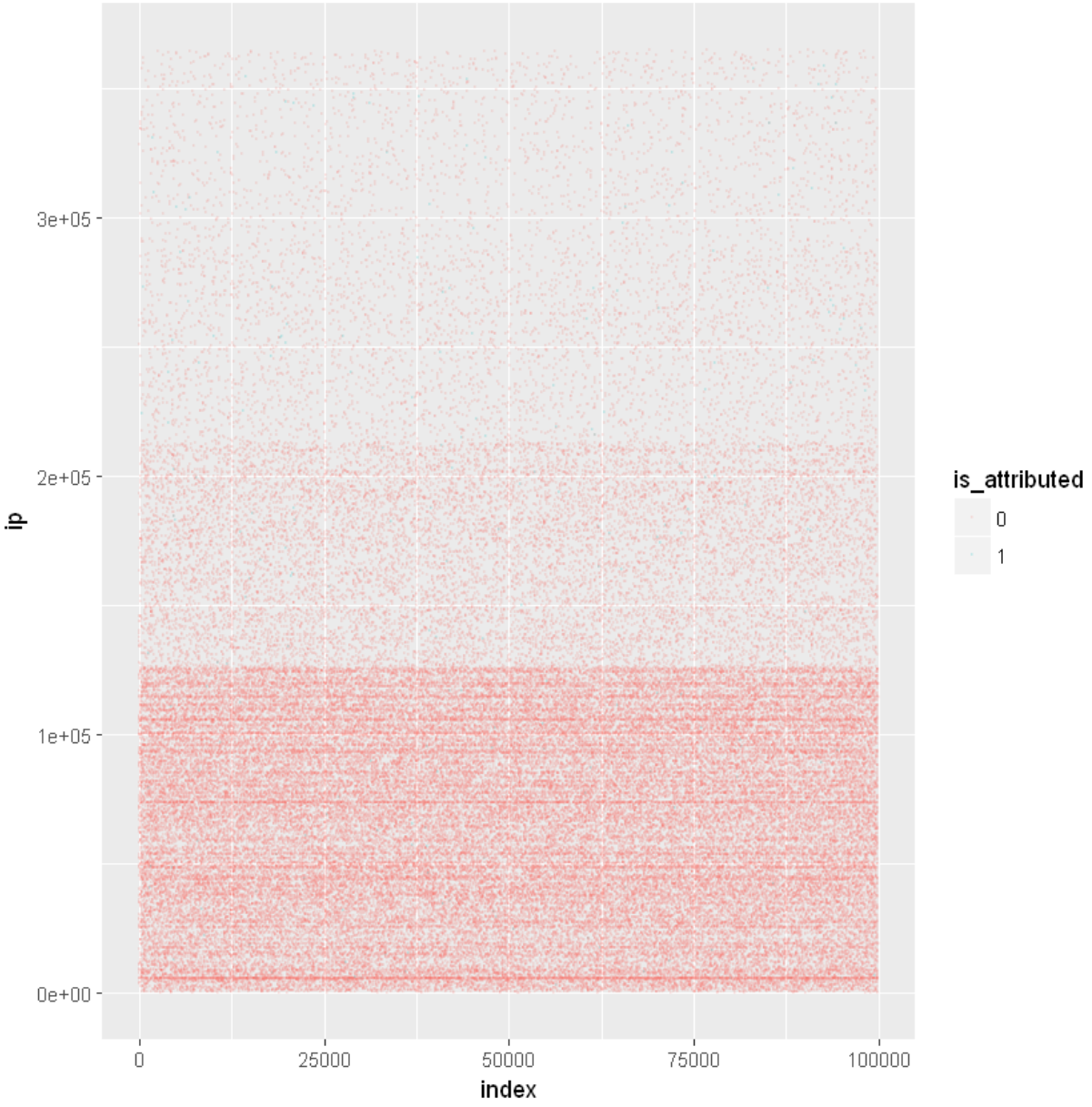


위의 plot으로 그래프를 작성한 것을 좀 더 ggplot을 이용하여 시각화를 수행해보자.

(1) ip 에 대해 아래와 같이 점 그래프를 작성해 보자.
(2) 색은 is_attribute로 구분되어 진다.

(3) alpha = 0.1로 지정했다. 점의 크기는 0.05이다.
(4) x축의 축이름은 : index로 지정했다.

In [6]:

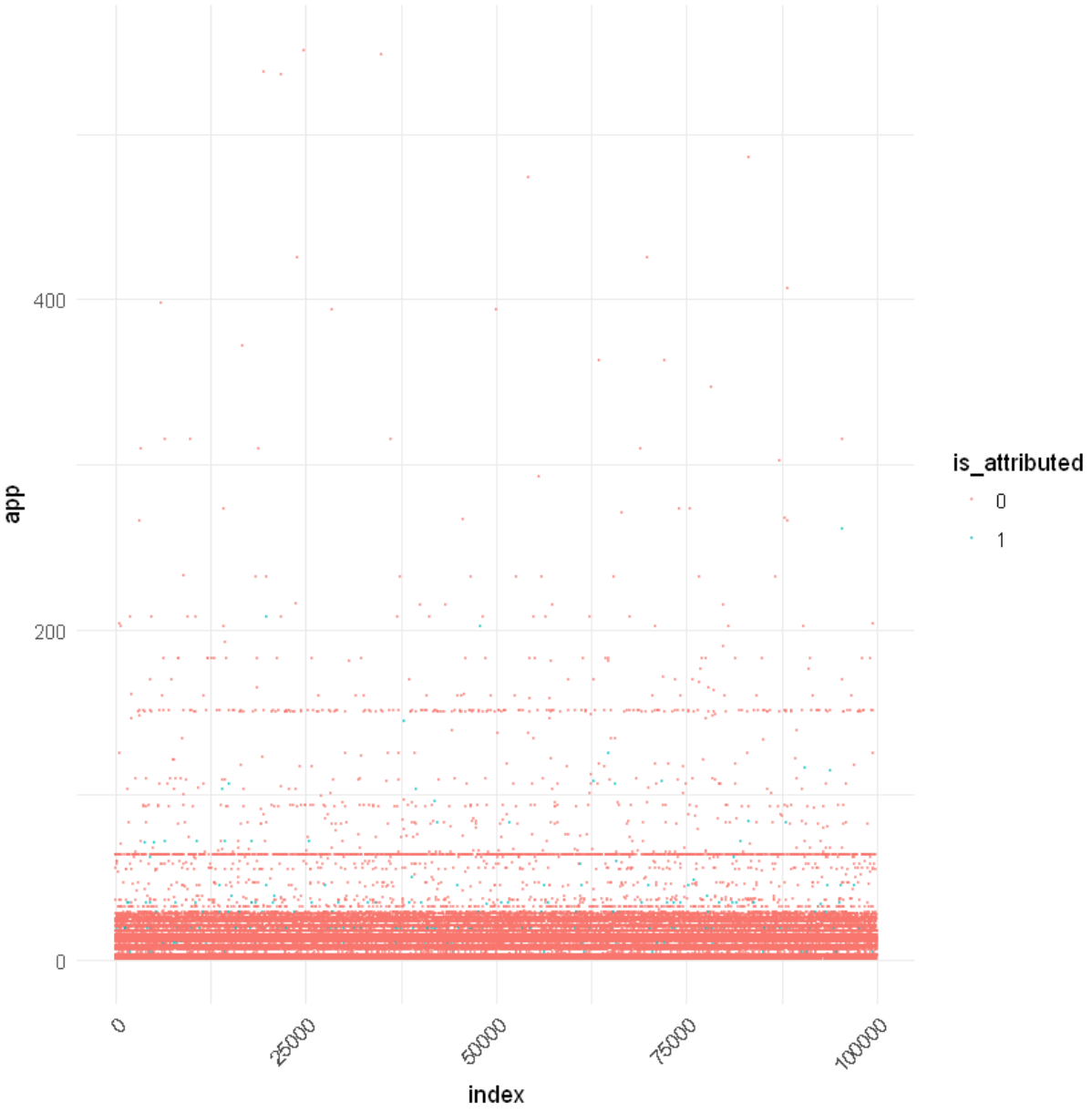


실습 과제1-2

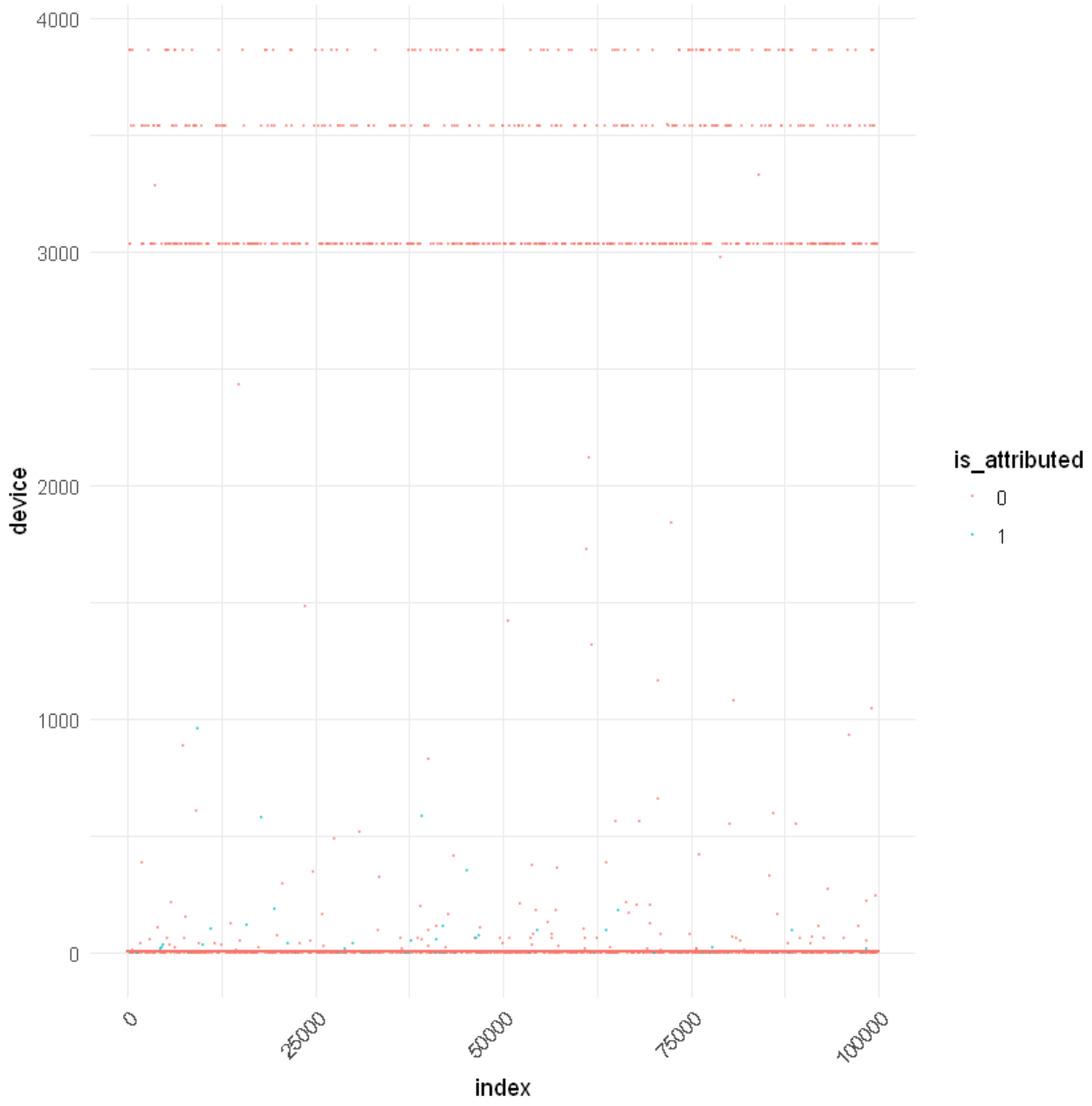
아래 그래프는 app, os, channel, device에 대한 내용을 시각화 해 보았다.

그래프의 생성된 조건은
(1) 범례가 있을 것,
(2) X축의 레이블의 값이 index로 변경하였다.
(3) theme를 이용하여 레이블의 값을 45로 기울게 하였고, 중앙 정렬하였다.

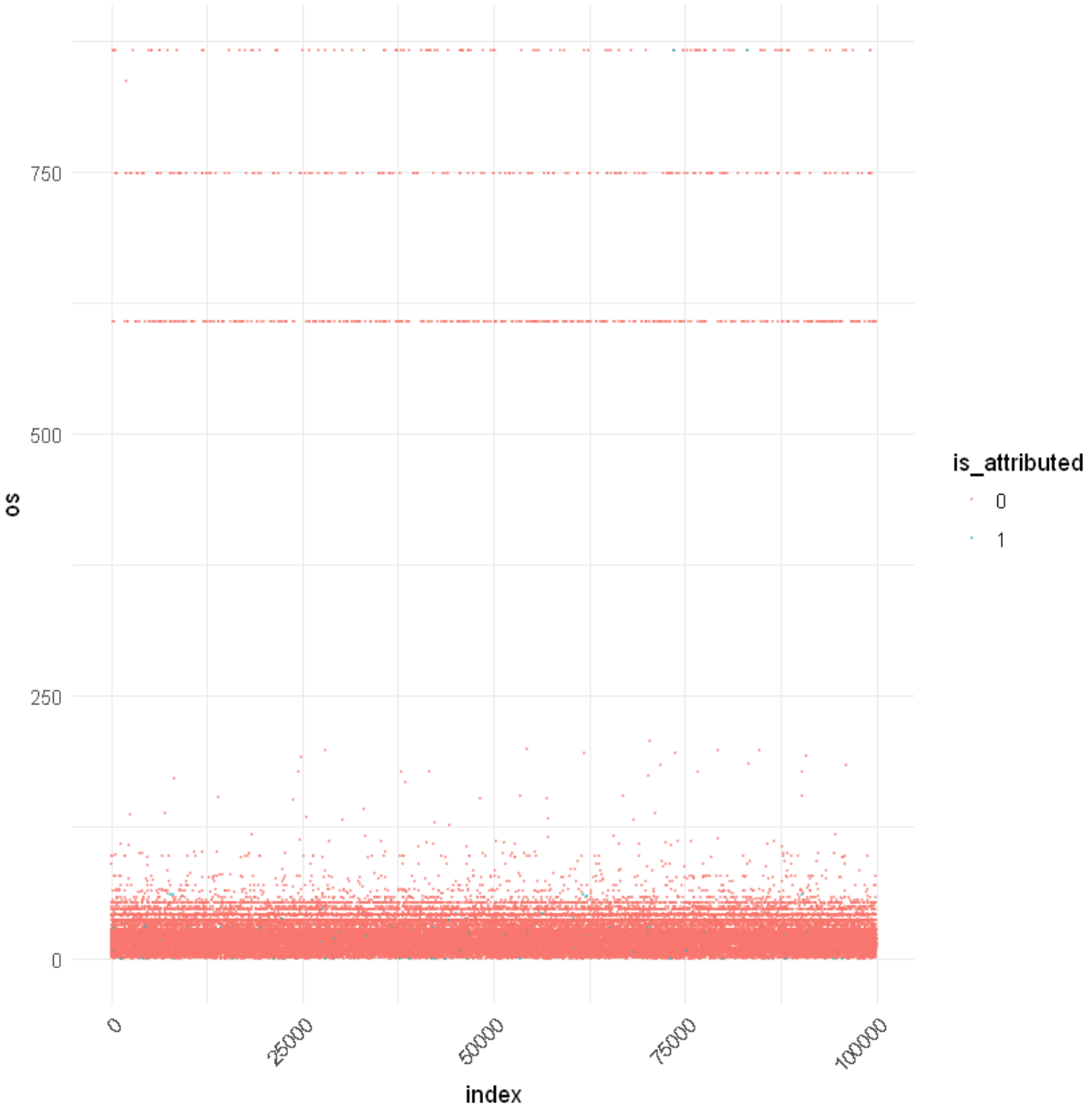
In [4]:


app
400
200
0
0
25000
50000
75000
100000
index
is_attributed
0
1

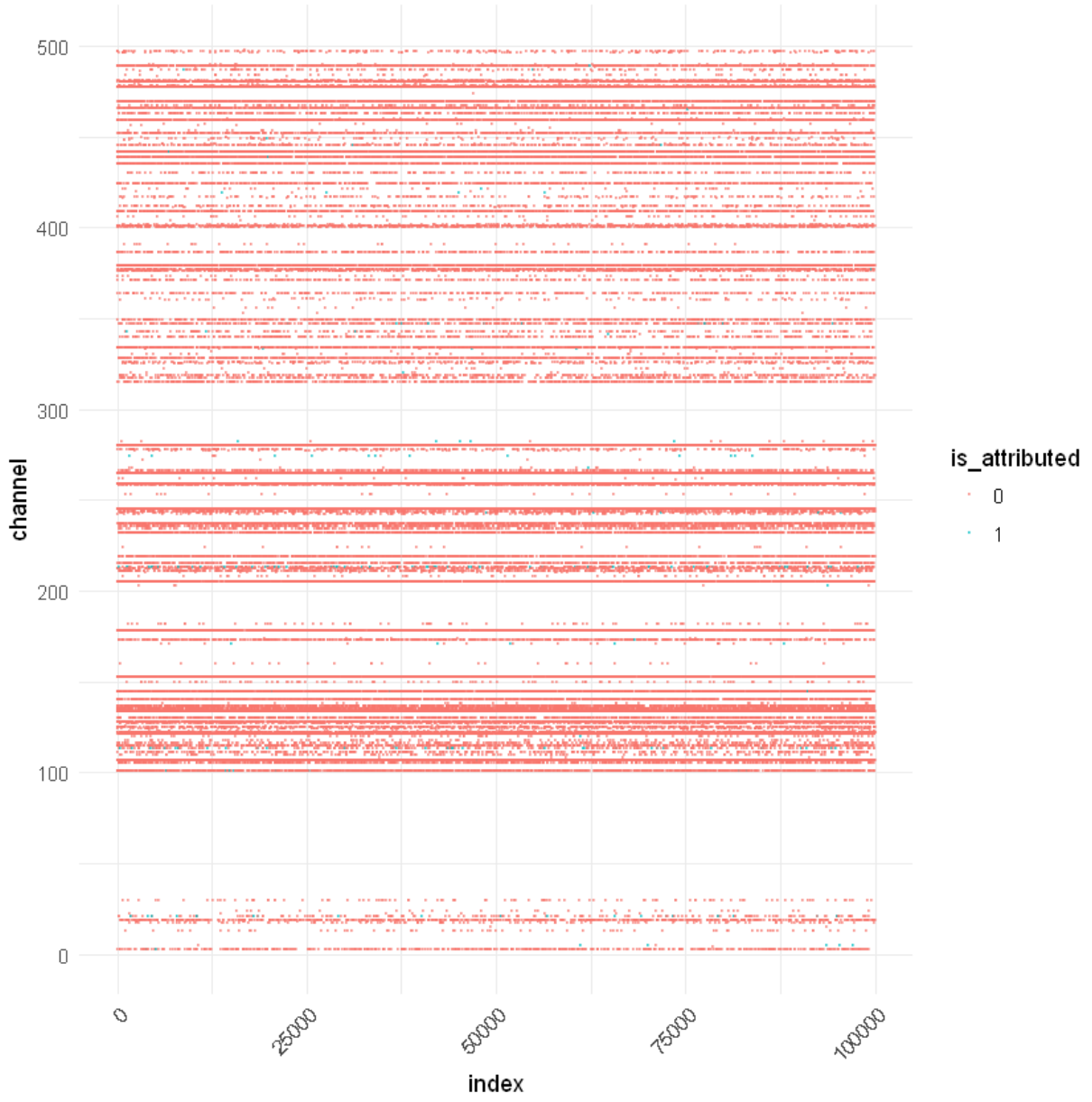
In [5]:


4000
3000
2000
1000
0
device
0
25000
50000
75000
100000
index
is_attributed
0
1

In [9]:


750
500
250
0
os
0
25000
50000
75000
100000
index
is_attributed
0
1

In [10]:



실습 과제1-3

위의 내용을 아래의 그래프 처럼 하나의 공간에 나누어서 그래프에 표시할 수 없을까?

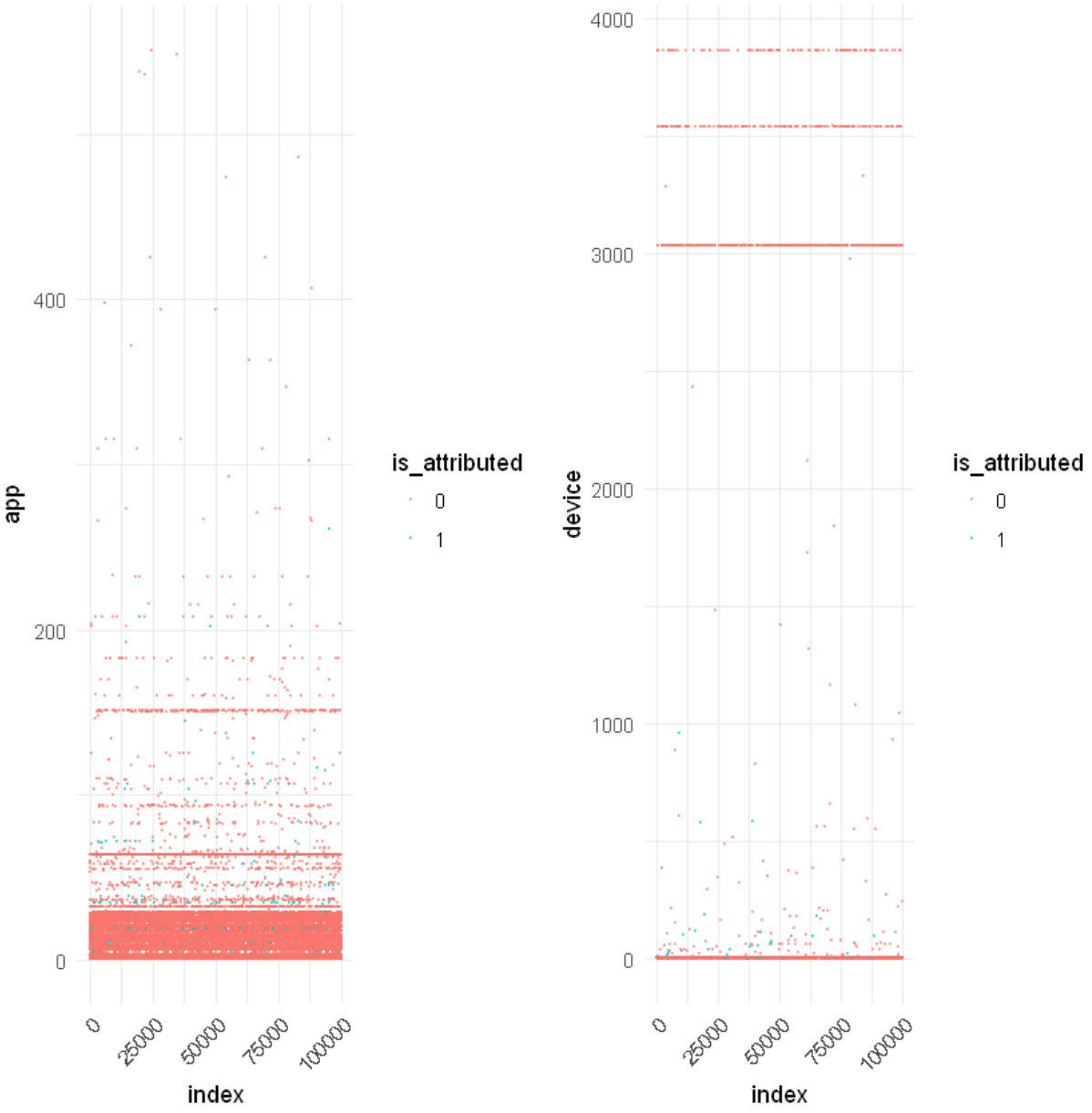
In [6]:
```
library(gridExtra)
```

```
Warning message:
"package 'gridExtra' was built under R version 3.4.4"
Attaching package: 'gridExtra'

The following object is masked from 'package:dplyr':

    combine
```

In [7]:
```
grid.arrange(pi1, pi2, ncol=2)
```



In [ ]:
```

```

실습과제 2. 훈련 데이터 나누기

우리는 5만개의 새로운 adTracking 훈련용 샘플 데이터를 확보했다.
이 데이터는 is_attributed 변수가 4만개, 1만개의 데이터를 각각 가지고 있다.
이중에 모델을 만들기 위해 1만개의 데이터 셋을 가지는 샘플을 만들어보자.

is_attributed 각각 7000(0), 3000(1) 이다. (sampling::strata) 함수 이용

In [9]:
```
library(sampling)
tr2 = read.csv("C:/data/ad_tracking/tr_5m_st4_1.csv")
names(tr2)
table(tr2$is_attributed)
```

```
Attaching package: 'sampling'

The following object is masked from 'package:caret':

    cluster
```

'X' 'ip' 'app' 'device' 'os' 'channel' 'click_time' 'attributed_time' 'is_attributed'

```
    0     1
40000 10000
```

비복원 추출이용 10000 개의 데이터를 만들기

In [ ]:
```

```

In [11]:
```
table(trS1$is_attributed)
names(trS1); dim(trS1)
table()
```

```
   0    1
7000 3000
```

'ip' 'app' 'device' 'os' 'channel' 'click_time' 'attributed_time' 'is_attributed'

10000 8

```
Error in table(): nothing to tabulate
Traceback:

1. table()
2. stop("nothing to tabulate")
```

In [12]:
```
write.csv(trS1, "C:/data/ad_tracking/tr_1m.csv", row.names=F)
list.files("C:/data/ad_tracking/")
```

'dfDat0413.RDS' 'img' 'img1.png' 'sample_submission.csv' 'sample_submission.csv.zip' 'sub_10m.csv' 'test.csv' 'test.csv.zip' 'test_10m.csv' 'testDat0413.RDS' 'tr_1m.csv' 'tr_5m_st4_1.csv' 'train.csv' 'train.csv.zip' 'train_sample.csv' 'train_sample.csv.zip' 'trainp1.csv' 'X_testmod_0414.RDS'

In [ ]:
```

```