De-duplication of outcomes

The problem

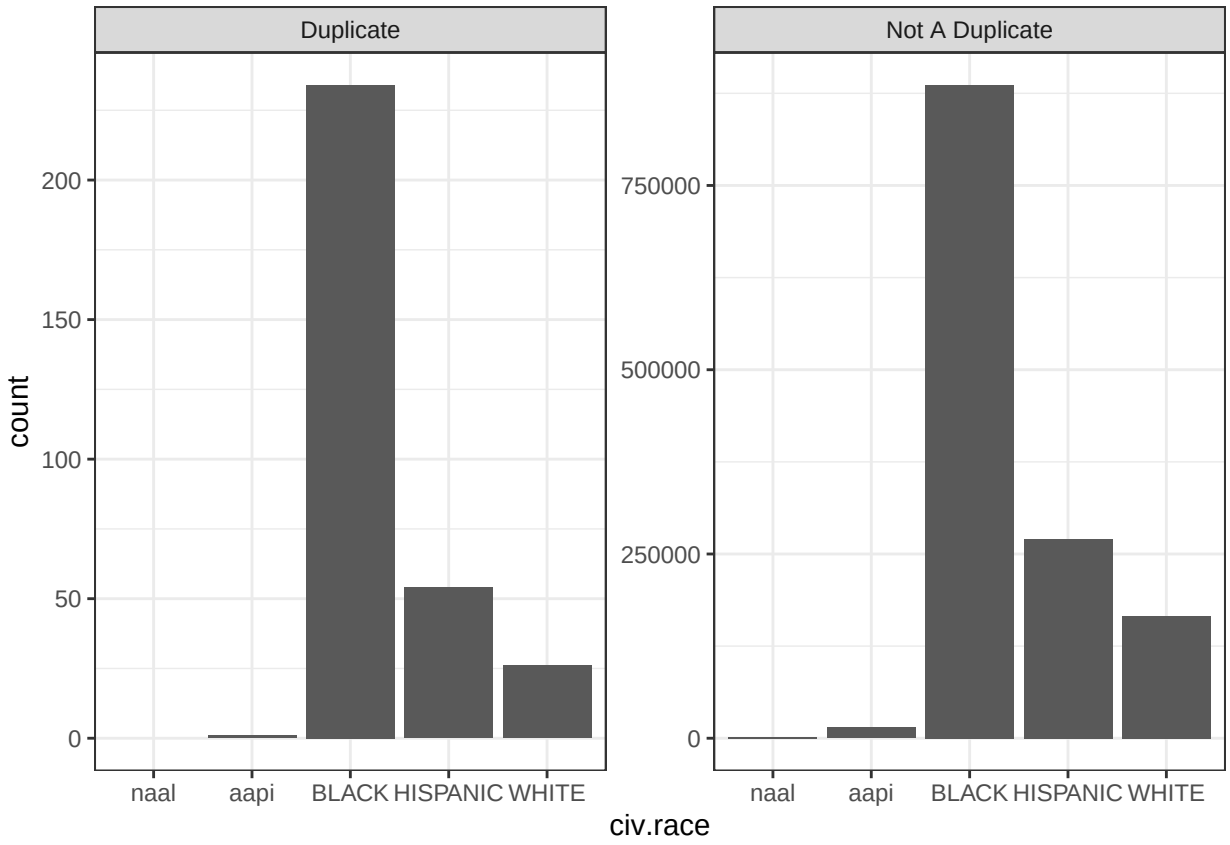
There are overlapping shifts and some police activity (e.g. stops) occurred during the overlap. It's unclear during which shift the activity **actually** occurred. Using the latitude and longitude from the police activity, I will attempt to discern what **beat** the activity occurred in and attempt to match that with the **beat where the shift assignment took place**.

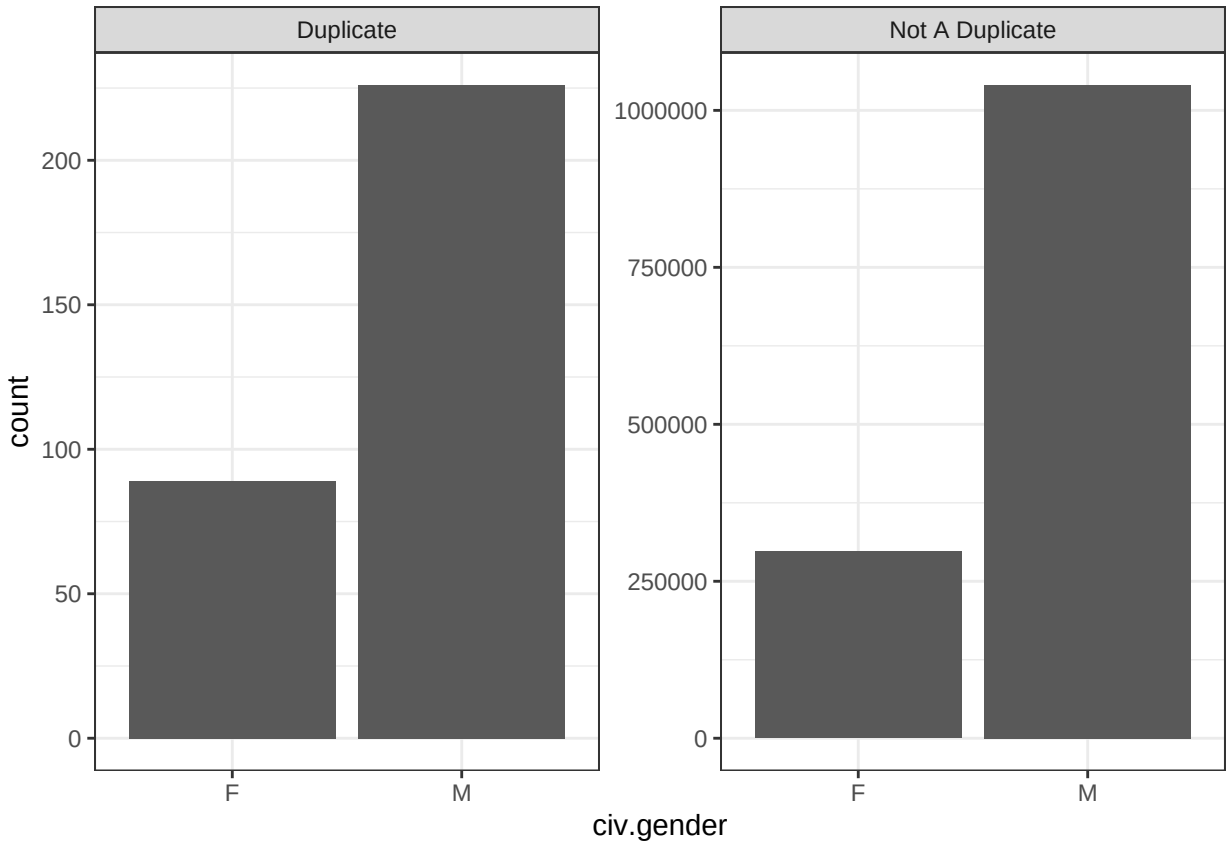
This method does not really work to clear up all duplicates, unfortunately. Some overlapping shifts were assigned to the same beat. Some police activity is missing a longitude and latitude. Other times the location of the police activity did not occur in any assigned beats as indicated in the shift assignment so a judgment call would have to be made as to which beat (and subsequently which shift assignment) the police activity actually occurred in. The distance from the police activity and each of the assigned beats could be used to discern during which shift the activity took place. However not all assigned beats correspond to a physical location. In the end, there is no foolproof method, and the very small number of affected cases does not warrant further effort for the time being.

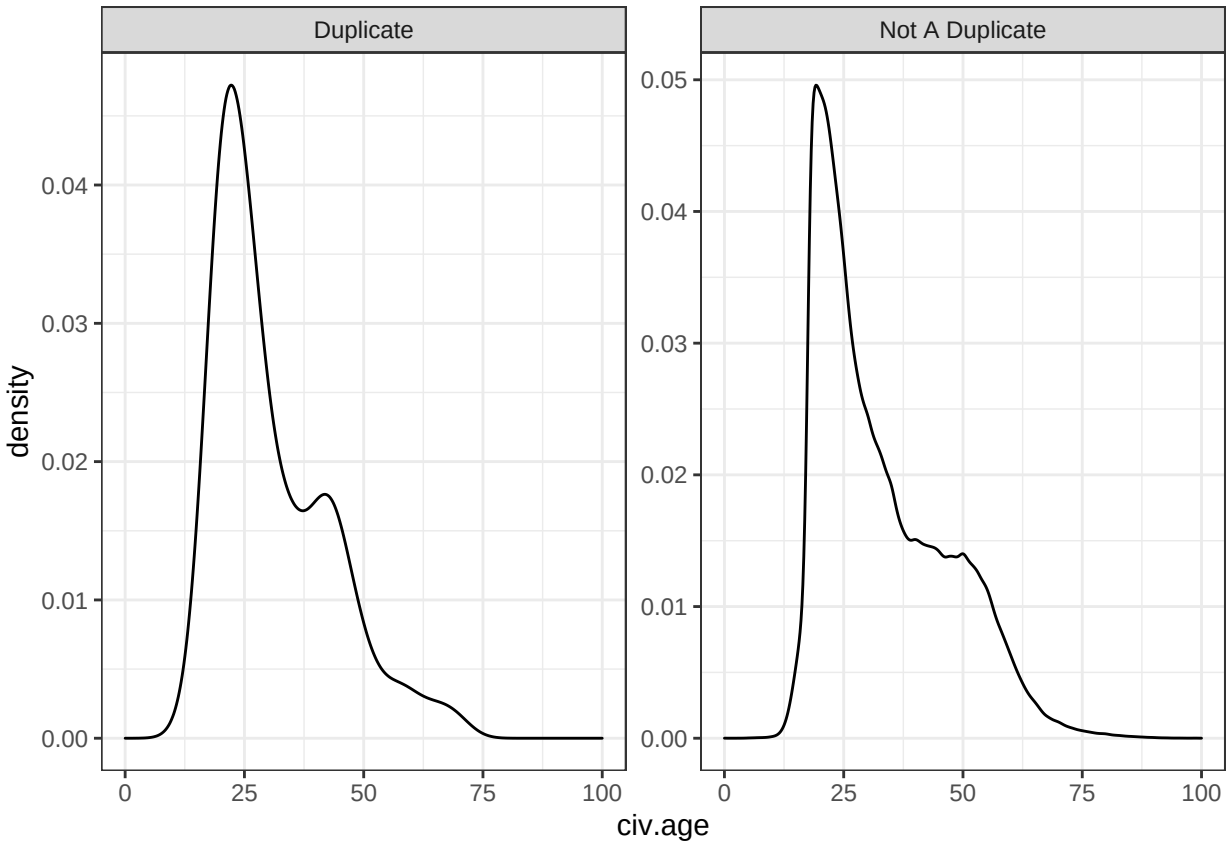
- Total number of duplicate entries: 630
  - Number of entries which are missing coordinates: 88
  - Number of entries with the same beat assignment: 334
  - Number of entries where a **match can be identified**: 68. Notice how few cases can be de-duplicated.
  - Number of entries where a **match cannot be identified**: 140. More work would have to be done to determine if these cases could be de-duplicated. Even then, it's a small number of cases.

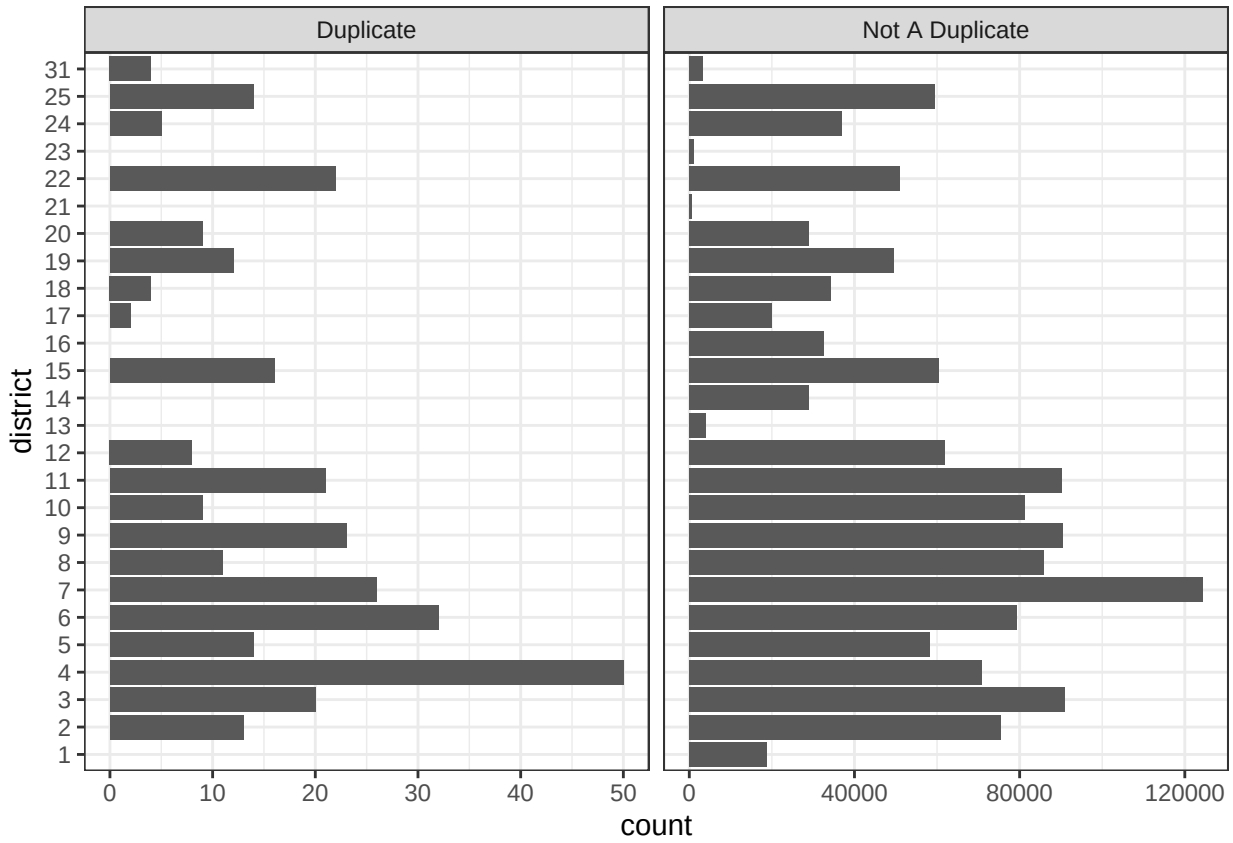
Do duplicates look like non-duplicates?

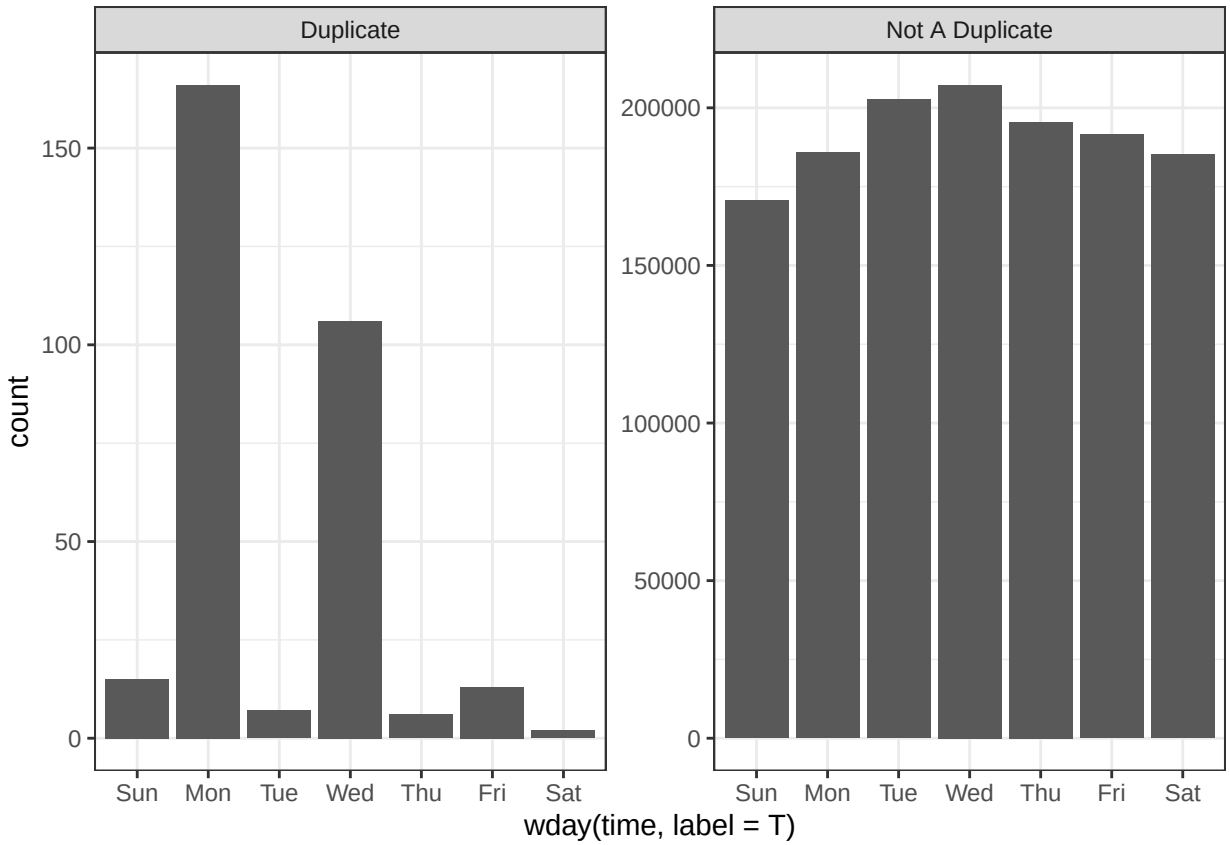
- Race looks the same.
- Gender looks the same.
- Age looks the same.
- District looks different, but there aren't many duplicates so I am not too concerned.
- Most duplicates happen during Monday and Wednesday. This is odd.
- Duplicates have a weird bimodal distribution as to when they appear the most.
- Contact type does not exactly match, but it is not terribly far off.

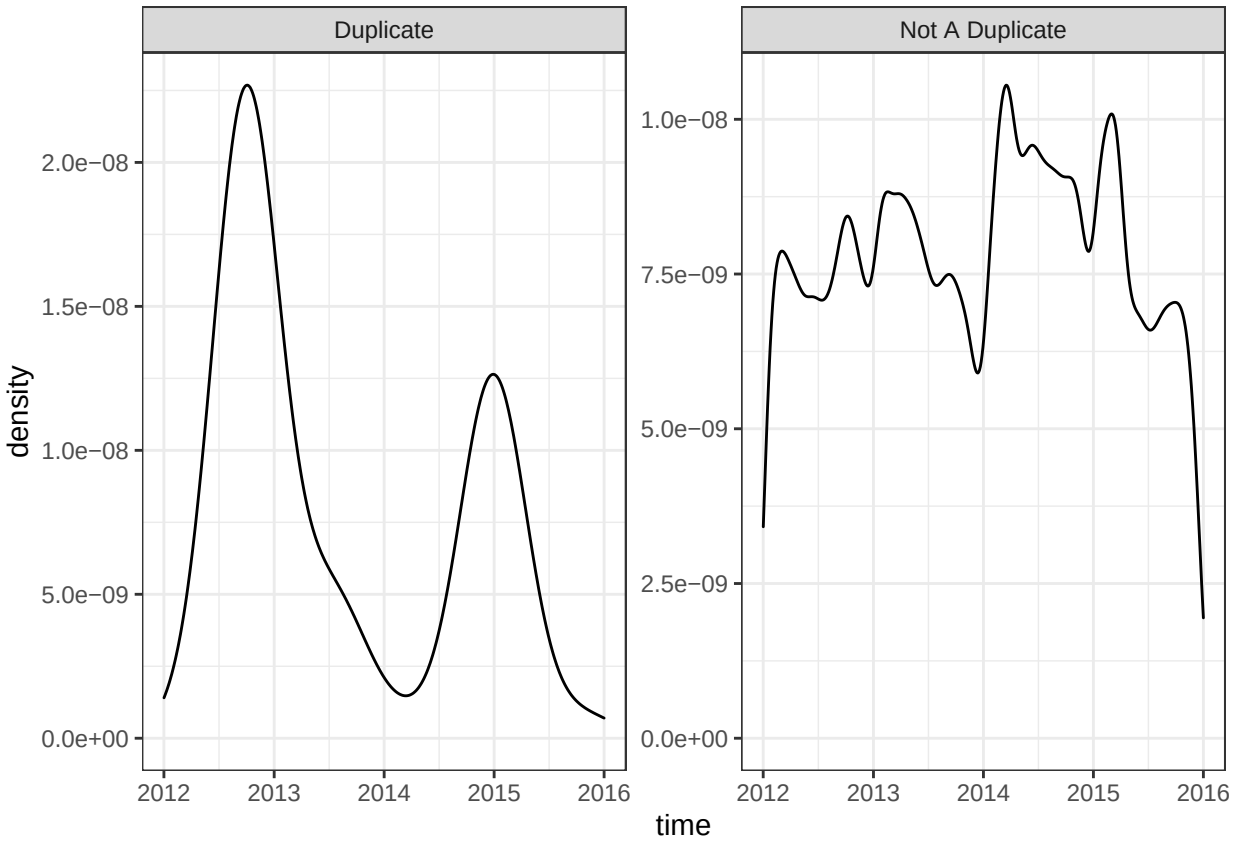

Duplicate
Not A Duplicate
count
200
150
100
50
0
750000
500000
250000
0
naal
aapi
BLACK
HISPANIC
WHITE
civ.race

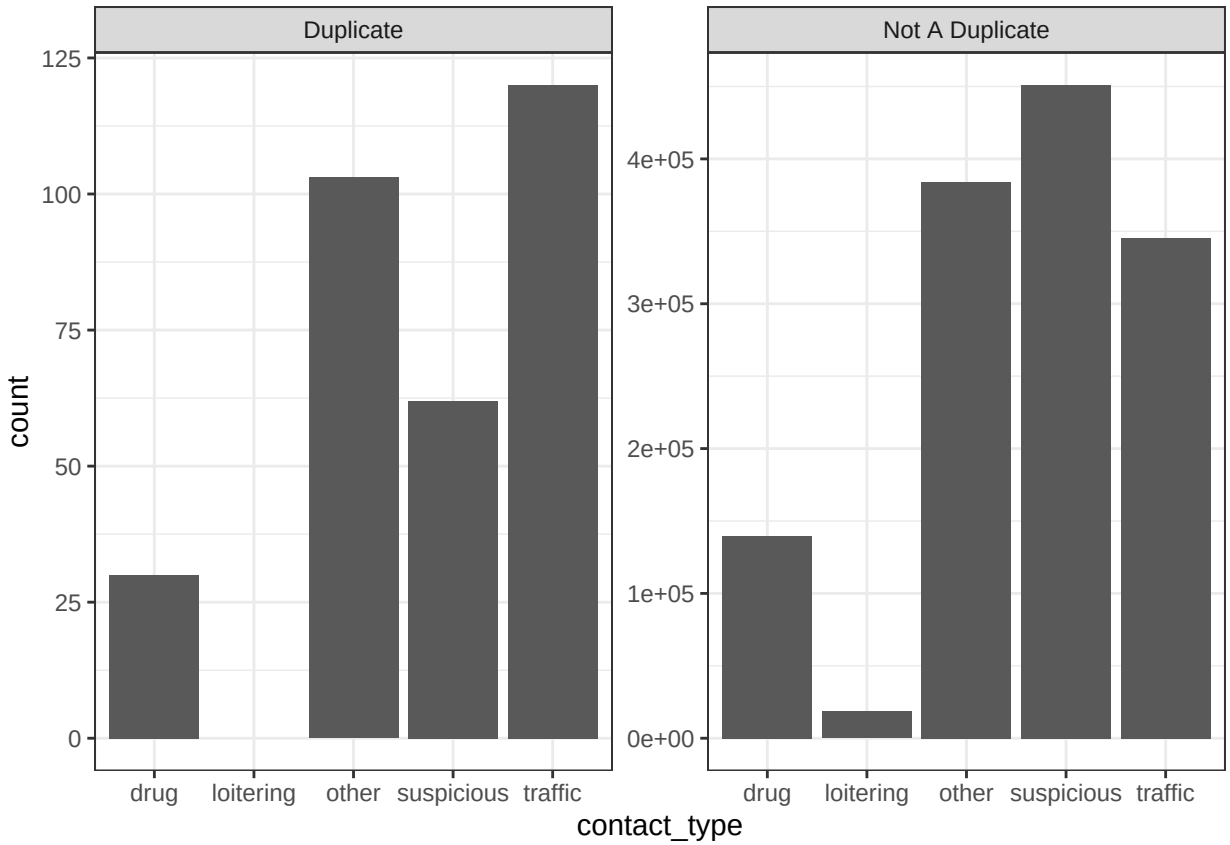

Duplicate
Not A Duplicate
count
civ.gender
0
50
100
150
200
250000
500000
750000
1000000
F
M


Duplicate
Not A Duplicate
density
civ.age
0.00
0.01
0.02
0.03
0.04
0.05
0
25
50
75
100


Duplicate
Not A Duplicate
district
31
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
10
20
30
40
50
0
40000
80000
120000
count


Duplicate
Not A Duplicate
count
150
100
50
0
200000
150000
100000
50000
0
Sun
Mon
Tue
Wed
Thu
Fri
Sat
wday(time, label = T)


Duplicate
Not A Duplicate
density
time
2.0e−08
1.5e−08
1.0e−08
5.0e−09
0.0e+00
7.5e−09
2.5e−09
2012
2013
2014
2015
2016


Duplicate
Not A Duplicate
count
125
100
75
50
25
0
4e+05
3e+05
2e+05
1e+05
0e+00
drug
loitering
other
suspicious
traffic
contact_type