Impute missing start and end times for work assignments

The logic and assumptions behind imputation

The key assumption being made is that officer-work assignments with a missing start time and end time will have the same start and end times as other **similar** officer-work assignments. If there was a systematic reason for the officer-work assignments to be missing their start and end times though, this assumption is not valid, and the imputation will not be valid.

Some of the missing officer-work assignments will be unique (i.e. there are not other similar officer work-assignments) in which case no values can be imputed for that officer-work assignment.

Similarity of officer-work assignments is systematically varied (e.g shift timing and beat **or** shift timing, beat, and month **or** shift timing, beat, month, and day of week **or** officer, beat, and shift timing), and the most common start and end times for officer-work assignments within each of these groupings is found.

As an example, officer 16811 was assigned to beat 110 on October 1, 2012 (Monday) during the 2nd shift. However, this officer-work assignment is missing its start and end times. So we look at all other officer-work assignments which took place in this beat during this shift timing and find the most common start/end times for those officer-work assignments. The most common start/end time (5:30am - 2:30pm) would then be the value given to this officer-work assignment missing its start/end time. Of course, one could argue one should only look at officer-work assignments which took place in this beat during this shift timing in the *month of October*. One might also argue you should only look at this specific officer and observe their start/end times when they work this beat during this specific shift timing. One can imagine many different ways to conceptualize **similar** sets of officer-work assignments to compare to.

What I will attempt to do here is try and disambiguate between all the different imputation possibilities.

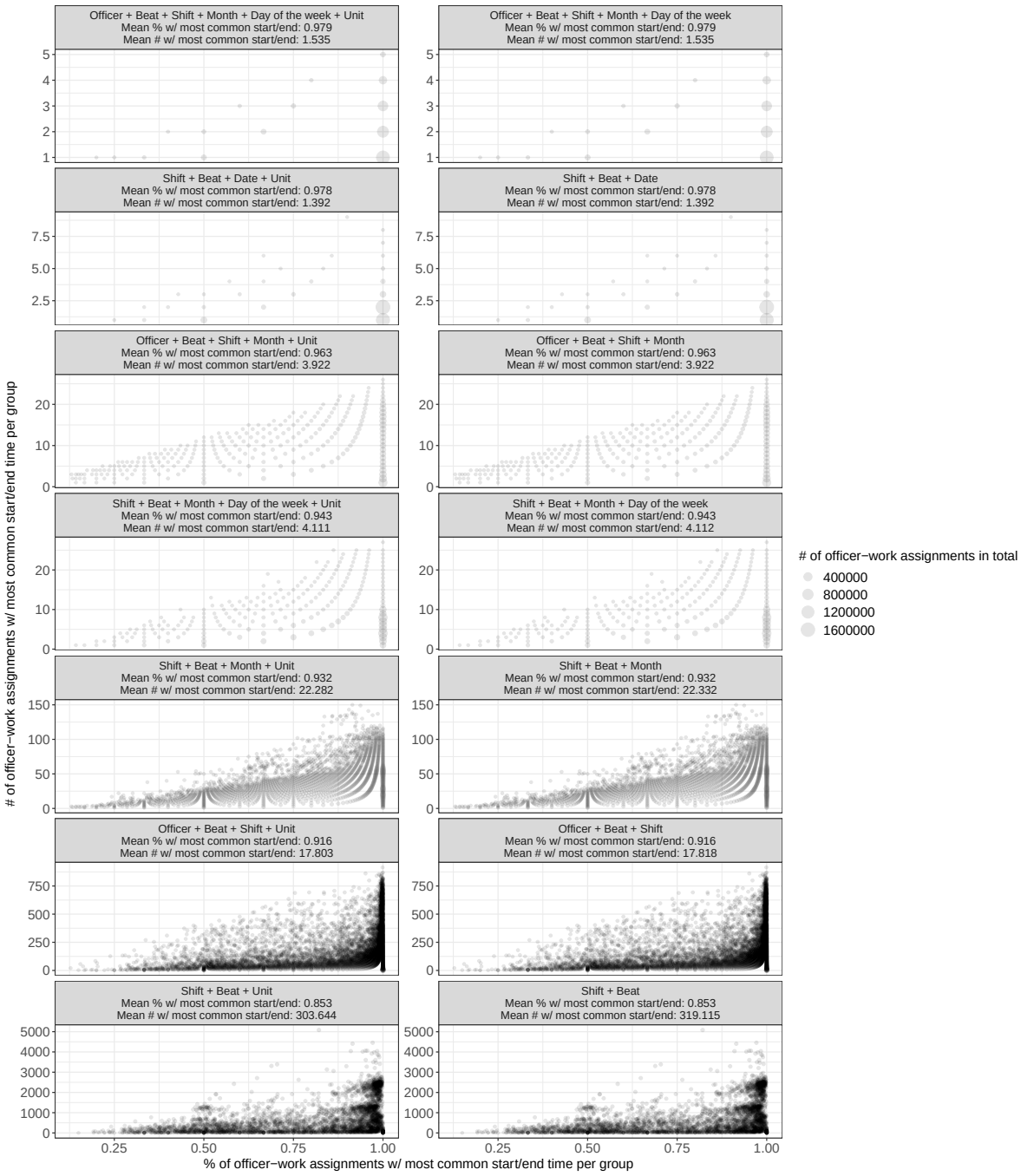
Evaluating different similiarity sets

Graphed below are 14 different ways I thought of to group together officer-work assignments. For each grouping, I visualize how well the imputation process works, in some sense. Using Shift + Beat as as an example, I will explain how to interpret the graph. First, I put every officer-work assignment into a group based on the beat and the shift timing. I.e., any officer-work assignment assigned to the same beat during the same shift timing is a part of the same group. Within each of these groups (based on the shift timing + beat), I found the most common start/end time for all officer-work assignments. Next, by assuming all officer-work assignments in a specific grouping had the most common start/end time, I calculated what percentage of the time the assumption is correct (x-axis). Then, I looked at how many officer-work assignments had the most common start/end time (y-axis). Finally, I combine groups if they have the exact same percentage and number of officer-work assignments with the most common start/end time (size of the circles). This last step is mostly to keep the graphic somewhat clean and not have so many overlapping dots.

As an example, look at shift + beat. Within shift + beat groups on average, 85% of officer-work assignments (or 319 officer-work assignments) have the most common start/end time. A specific dot represents a set of shift + beat groups where Y number of officer-work assignments (or X% of all officer-work assignments in those groups) share the most common start/end time. The size of the circle represents the total number of officer-work assignments in the set of groups (regardless of if they share the most common start/end time).

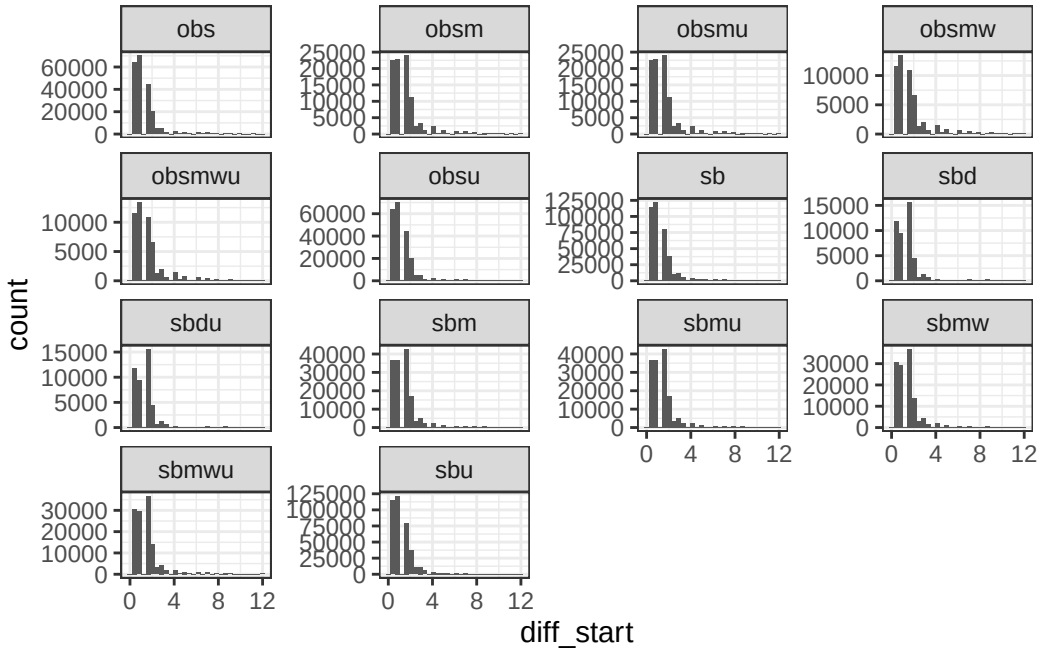
The main takeaway from the graph is the fact that as one increases the exclusivity of the groups, the higher the percentage of officer-work assignments which share the most common start/end time (meaning if we were to impute a missing value, we would be highly confident the start/end time would be correct). However, this mechanically means there is a smaller absolute number of officer-work assignments which we are comparing against, and in the worst case there may only be a group of 1 in which case it would not be helpful for imputation at all.

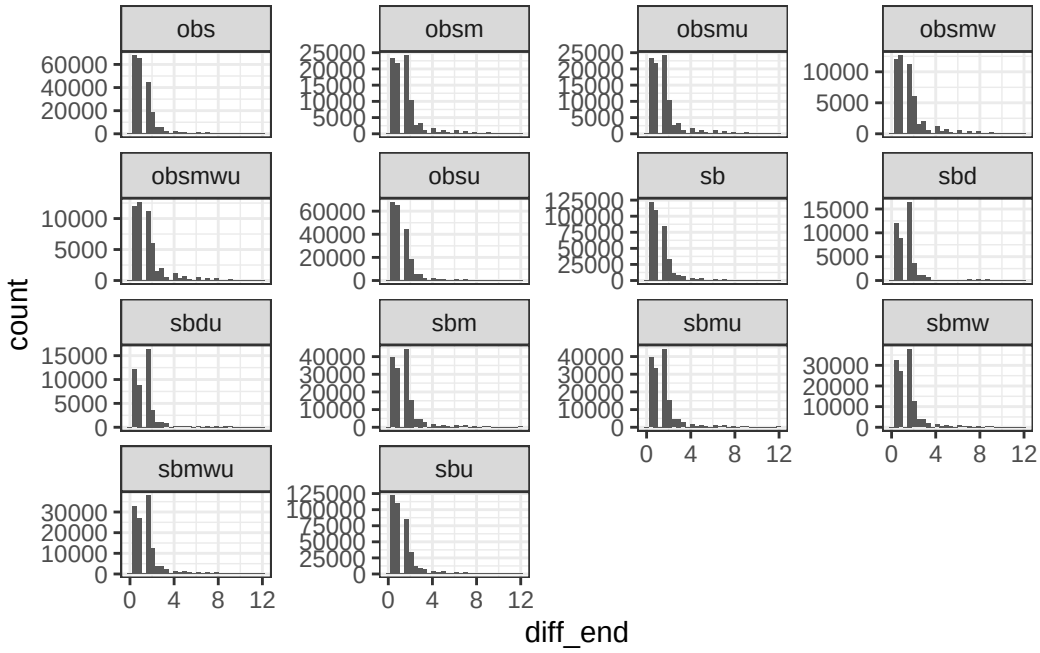
Also, in an absolute sense, we have can have some confidence that even in the worst case, we are still about 85% likely to be correct in our imputation (see Shift + Beat).



Using the above process of imputation, how far off the mark are we when we impute the wrong start/end time? The amount we are off the mark does not vary greatly with how we define similarity sets. On average, the imputation is off by about 1 hour to 1.5 hours with 75% of values between 0 and 2 hours. It is hard to say in an absolute sense how good/bad this is. It

does not seem that bad?





group	min_start	q1_start	median_start	mean_start	q3_start	max_start	min_end	q1_end	median_end	mean_end	q3_end	max_end
obs	0.03	0.5	1.0	1.36	1.5	12	0.02	0.5	1.0	1.39	1.5	12
obsm	0.03	1.0	1.5	1.65	2.0	12	0.02	1.0	1.5	1.67	2.0	12
obsmu	0.03	1.0	1.5	1.65	2.0	12	0.02	1.0	1.5	1.67	2.0	12
obsmw	0.15	1.0	1.5	1.74	2.0	12	0.02	1.0	1.5	1.76	2.0	12
obsmwu	0.15	1.0	1.5	1.74	2.0	12	0.02	1.0	1.5	1.76	2.0	12
obsu	0.03	0.5	1.0	1.36	1.5	12	0.02	0.5	1.0	1.39	1.5	12
sb	0.03	0.5	1.0	1.40	1.5	12	0.02	0.5	1.0	1.46	1.5	12
sbd	0.03	0.5	1.5	1.52	1.5	12	0.02	0.5	1.5	1.52	1.5	12
sbdu	0.03	0.5	1.5	1.52	1.5	12	0.02	0.5	1.5	1.52	1.5	12
sbm	0.03	1.0	1.5	1.61	2.0	12	0.02	0.5	1.5	1.62	1.5	12
sbmu	0.03	1.0	1.5	1.61	2.0	12	0.02	0.5	1.5	1.62	1.5	12
sbmw	0.03	1.0	1.5	1.60	1.5	12	0.02	0.5	1.5	1.61	1.5	12
sbmwu	0.03	1.0	1.5	1.60	1.5	12	0.02	0.5	1.5	1.61	1.5	12
sbu	0.03	0.5	1.0	1.40	1.5	12	0.02	0.5	1.0	1.46	1.5	12

Which method to use?

As a refresher, we have 48210 officer-work assignments with a missing start/end time (out of 3519518 officer-work assignments). For each of these missing officer-work assignments, we impute a start/end time using the most common start/end time based on the group that the officer-work assignment is a part of (where group is defined in 14 different ways). How often do these estimates align? Quite a lot, thankfully. Roughly, 77.7% of all missing officer-work assignments (which **can** be imputed) have the same estimated start/end time across all 14 methods.

The question now is should we try and impute the rest? If we do, what method should we use? One idea I had was:

- Select those estimates which come from those grouping definitions that have the highest percentage of officer-work assignments sharing in the most common start/end time. **NOTE**: Undecided on if we should include some sort of cutoff like the percentage has to be greater than 85%.

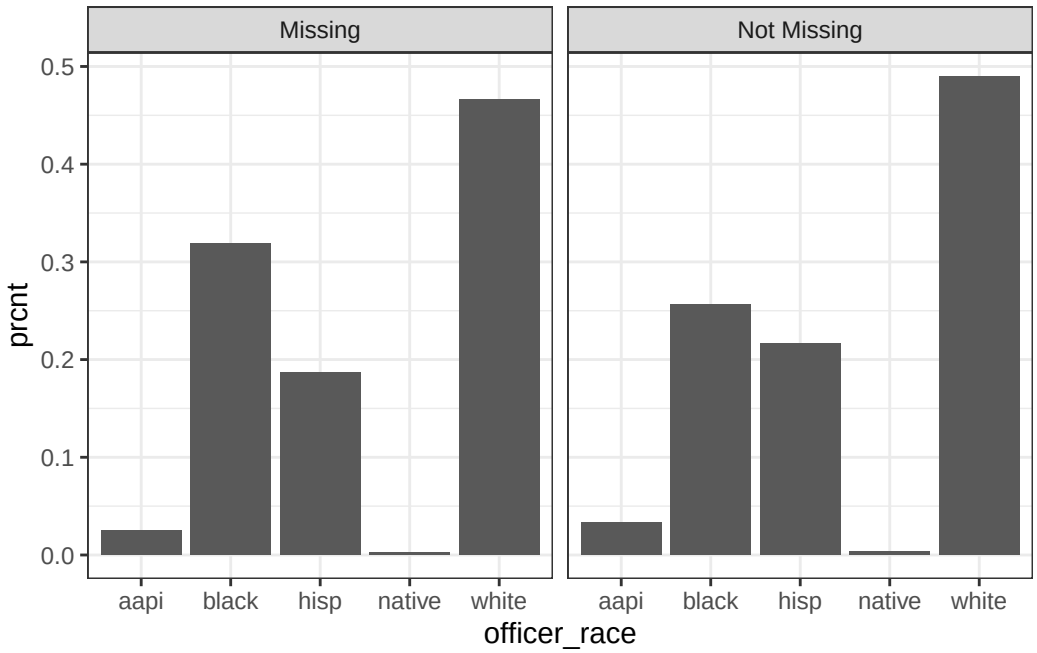
- If all estimates still do not match, keep those estimates from those grouping definitions with the highest absolute number of officer-work assignments that share in the most common start/end time.
- If all estimates still do no match, assert that some groupings are preferred over other groupings, all else equal, and use the estimates from that grouping.

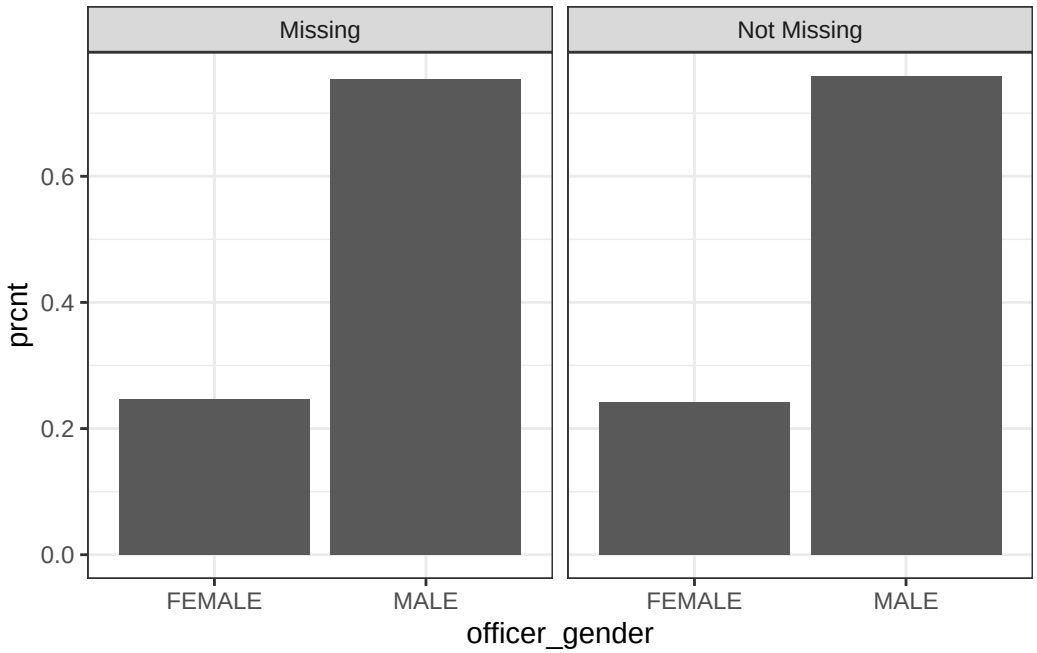
We could also take the minimum start time of the estimates and the maximum end time of the estimates to create our estimate. We also may decide not to impute these officer-work assignments.

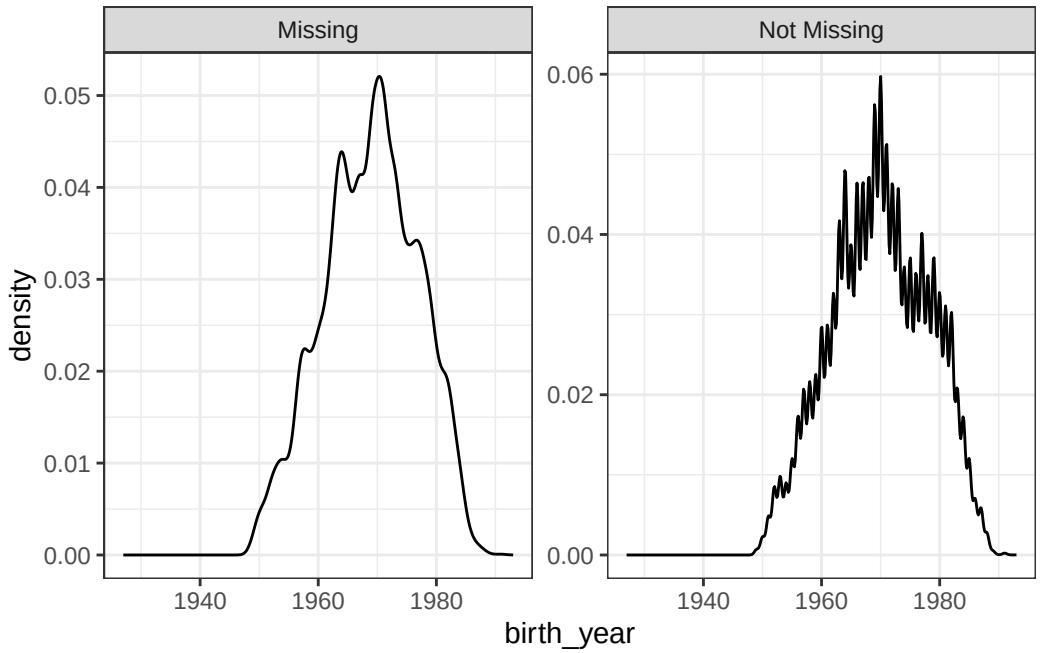
As a final note, there are 125 officer-work assignments which cannot be imputed. They have no other similar officer-work assignments regardless of the grouping definition.

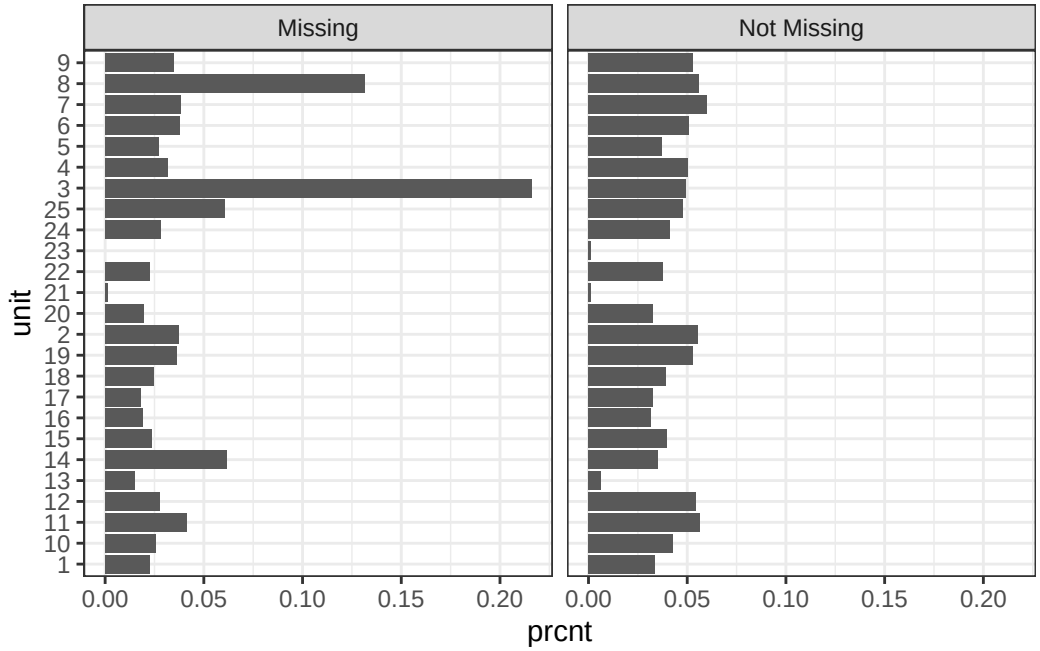
Do shift assignments with missing start/end times look like non-missing shift assignments?

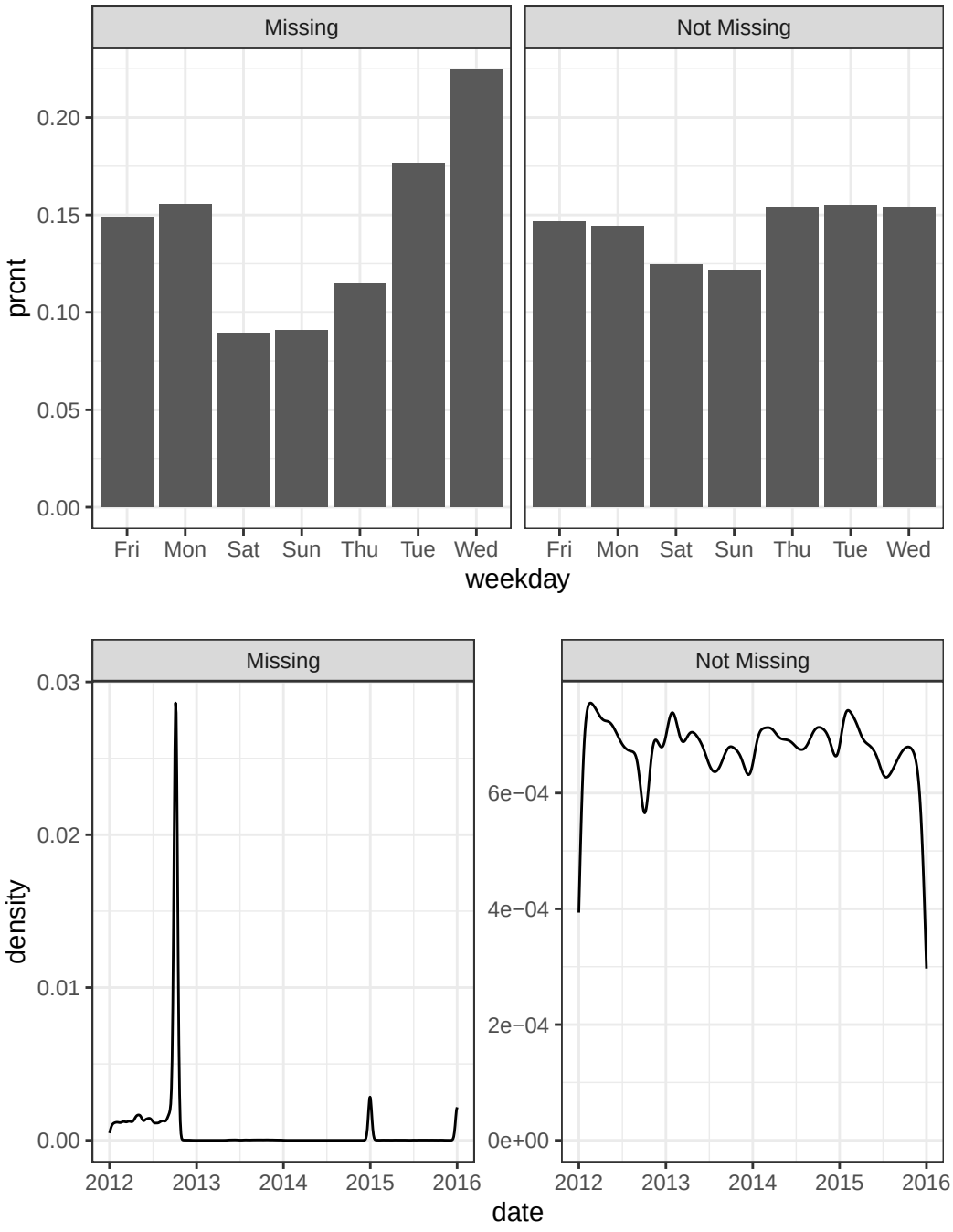
- Race largely looks the same. It would appear as if slightly more Black officers have missing officer-work assignments.
- Gender looks the same.
- Age looks the same.
- Spanish-speakers look the same.
- Disproportionately, units 3 and 8 have the most officer-work assignments with missing times.
- Weekday distributions don't completely align, but I am not concerned.
- Missing shift times occur almost entirely in 2012 which is an intriguing finding.
- Shift timing distribution looks roughly the same with the 1st shift being slightly under-represented.
- Rank distribution looks largely the same.

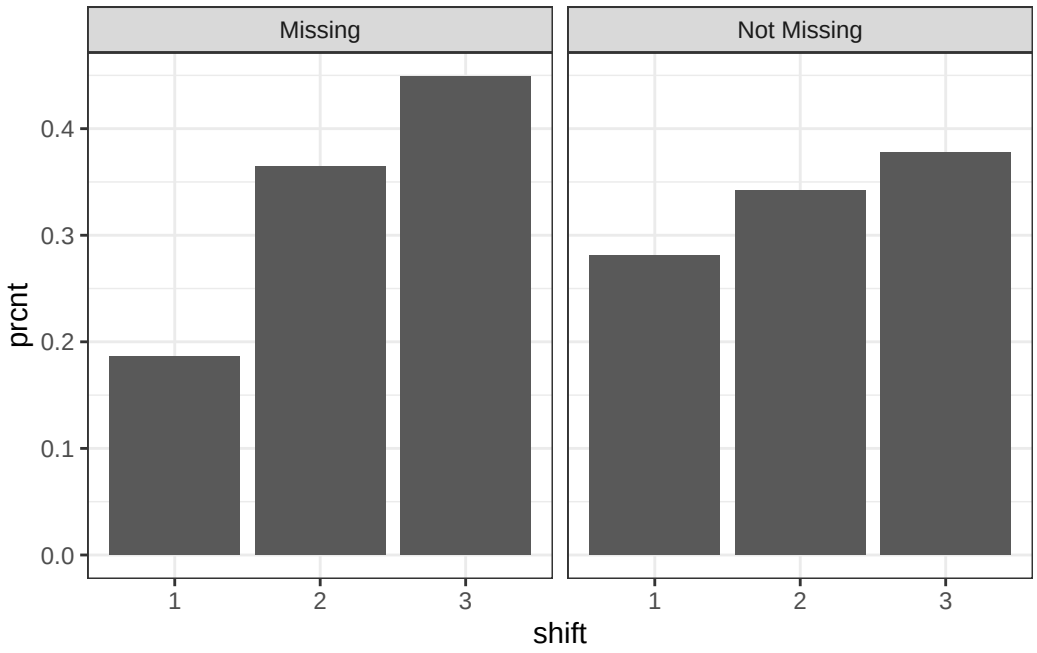

Missing
Not Missing
prcnt
0.5
0.4
0.3
0.2
0.1
0.0
aapi
black
hisp
native
white
officer_race

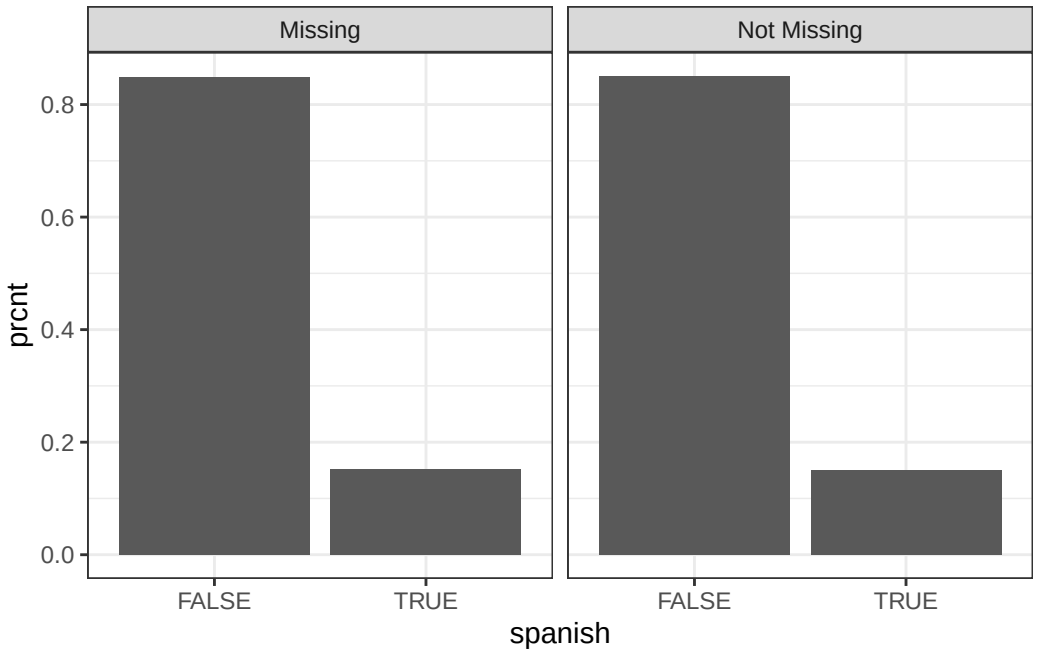

Missing
Not Missing
prcnt
0.6
0.4
0.2
0.0
FEMALE
MALE
officer_gender

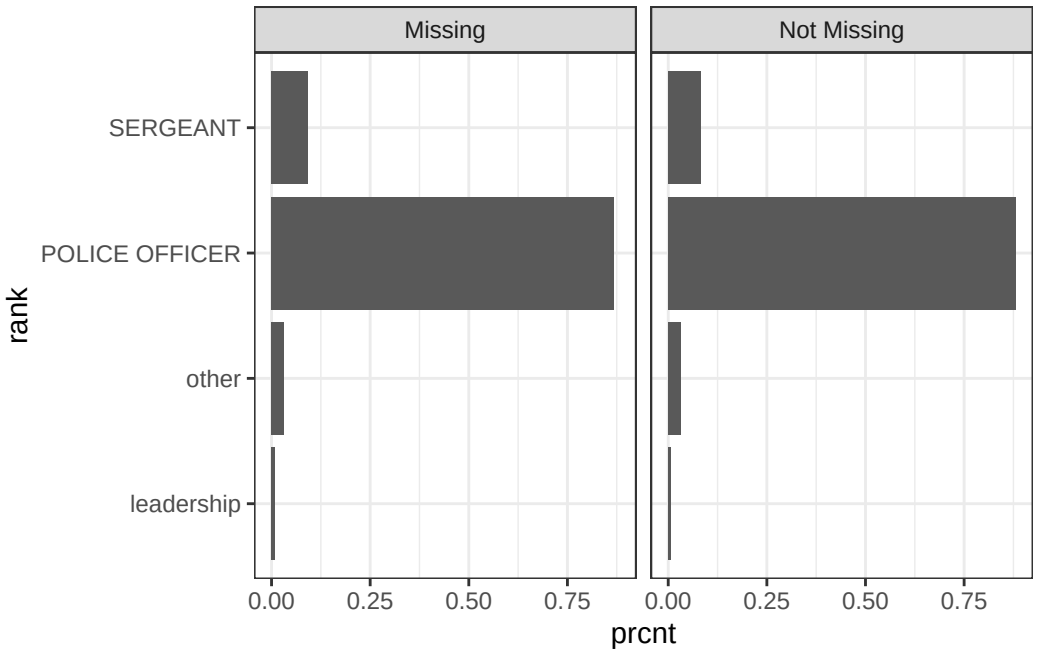

Missing
Not Missing
density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
1940
1960
1980
birth_year


Missing
Not Missing
unit
9
8
7
6
5
4
3
25
24
23
22
21
20
2
19
18
17
16
15
14
13
12
11
10
1
0.00
0.05
0.10
0.15
0.20
prcnt


Missing
Not Missing
prcnt
0.20
0.15
0.10
0.05
0.00
Fri
Mon
Sat
Sun
Thu
Tue
Wed
weekday
Missing
Not Missing
density
0.03
0.02
0.01
0.00
6e−04
4e−04
2e−04
0e+00
2012
2013
2014
2015
2016
date


Missing
Not Missing
prcnt
0.0
0.1
0.2
0.3
0.4
1
2
3
shift


Missing
Not Missing
prcnt
0.0
0.2
0.4
0.6
0.8
FALSE
TRUE
spanish


Missing
Not Missing
SERGEANT
POLICE OFFICER
other
leadership
rank
0.00
0.25
0.50
0.75
0.00
0.25
0.50
0.75
prcnt