Estimating differences in duplicated entries

None of the values which did not match between the cross-sectional and longitudinal data had to be de-duped. Stated more plainly, any observation which had different cross-section vs. longitudinal estimates did not have more than observation for a given year or year range in which the estimates were different.

Early (or later) means the value from the earlier (or later) year was selected in cases where we had multiple observations for the same year (or year range), and the values were different. **Same value** indicates there were multiple observations for the same year (or year range), and both values were the same. Only one needed to kept as a result. **Kept non-missing** indicates there multiple observations for the same year (or year range), and one of the values was missing so I simply kept the non-missing value. Finally **no duplicate** indicates the there were not multiple observations for the same year (or year range).

Var1	Freq
early	6,848
no duplicate	3,813,372
same value	477,376

Var1	Freq
early	0.1593449
no duplicate	88.7326775
same value	11.1079776

- **Number of rows**: 4297596
- **Number of non-duplicates (regardless of same values)**: 4290748
- **Percent of non-duplicates (regardless of same values)**: 99.8406551

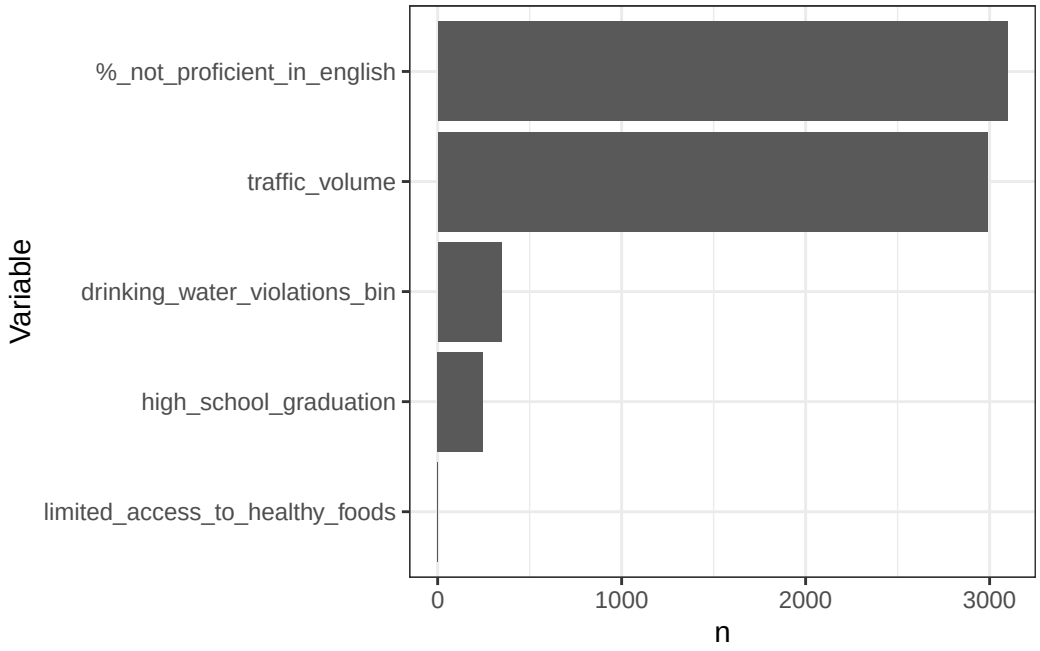
Table 1: Comparing deduped values across Cross Sectional and Longitudinal data

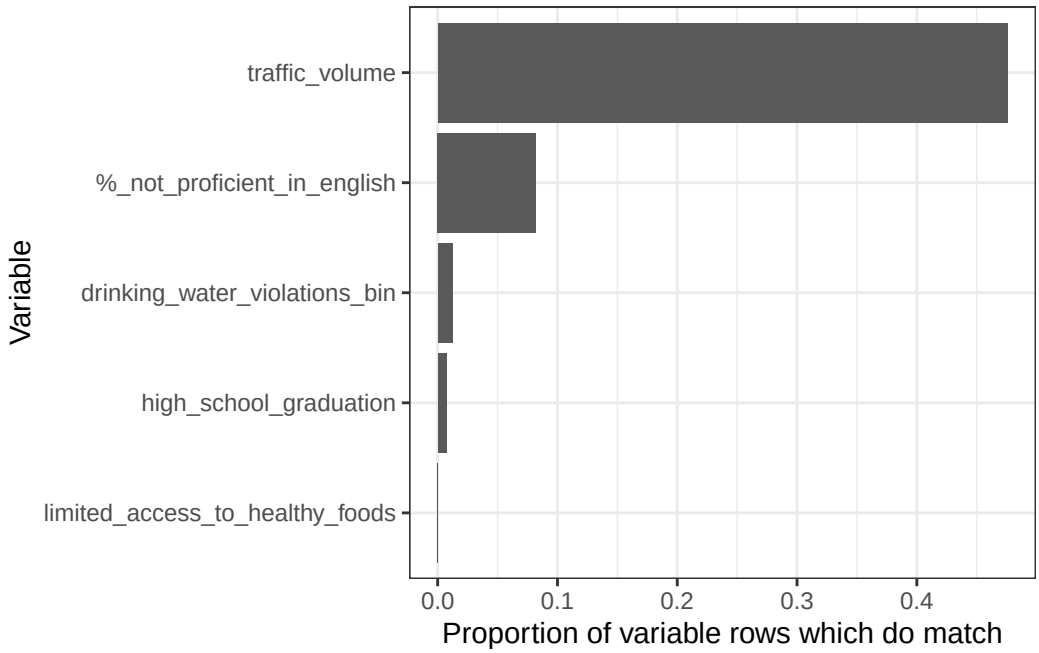
(a) Cross sectional

Var1	Var2	Freq
match	early	6848
no match	early	0
match	no duplicate	3554515
no match	no duplicate	258857
match	same value	471191
no match	same value	6185

(b) Longitudinal

Var1	Var2	Freq
match	early	6848
no match	early	0
match	no duplicate	3554515
no match	no duplicate	258857
match	same value	471191
no match	same value	6185

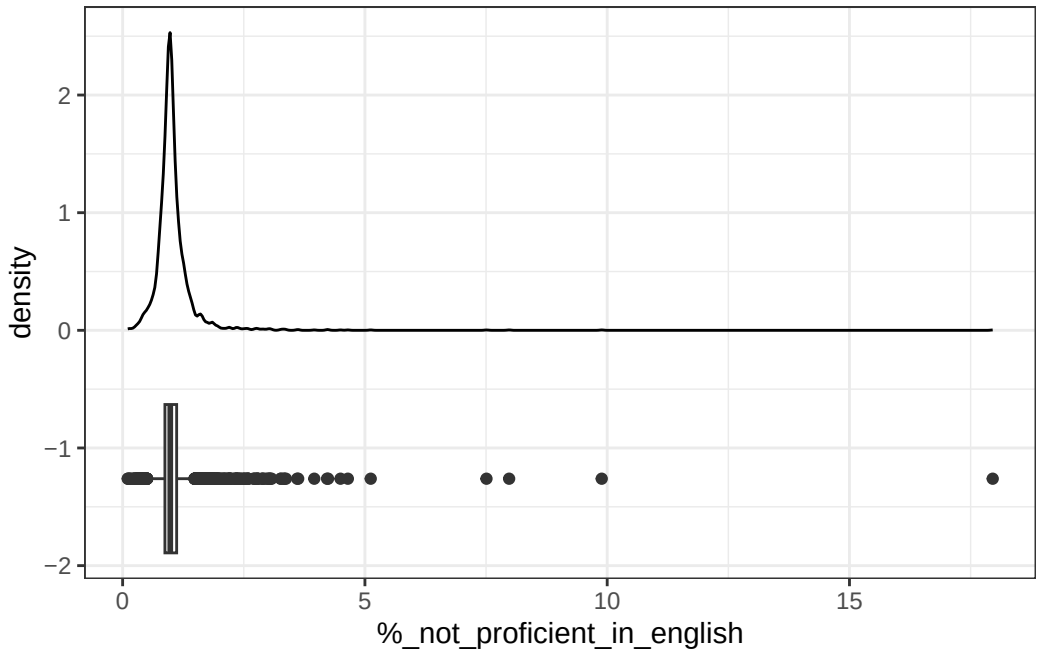





```
[[1]]
```

```
Warning: Removed 41 rows containing non-finite values (`stat_boxplot()`).
```

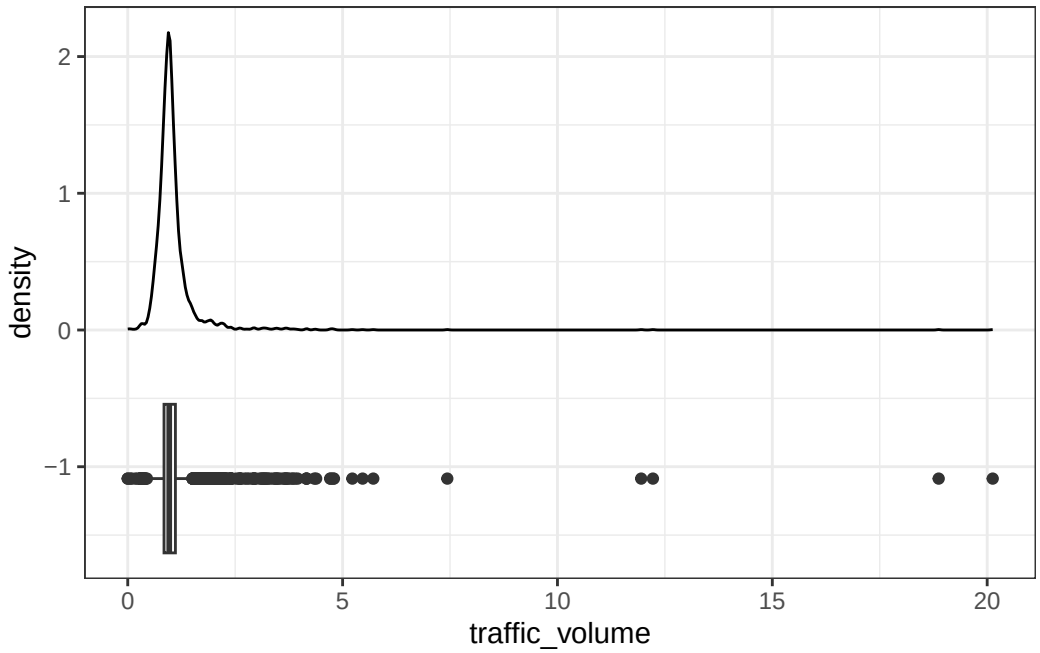
```
Warning: Removed 41 rows containing non-finite values (`stat_density()`).
```



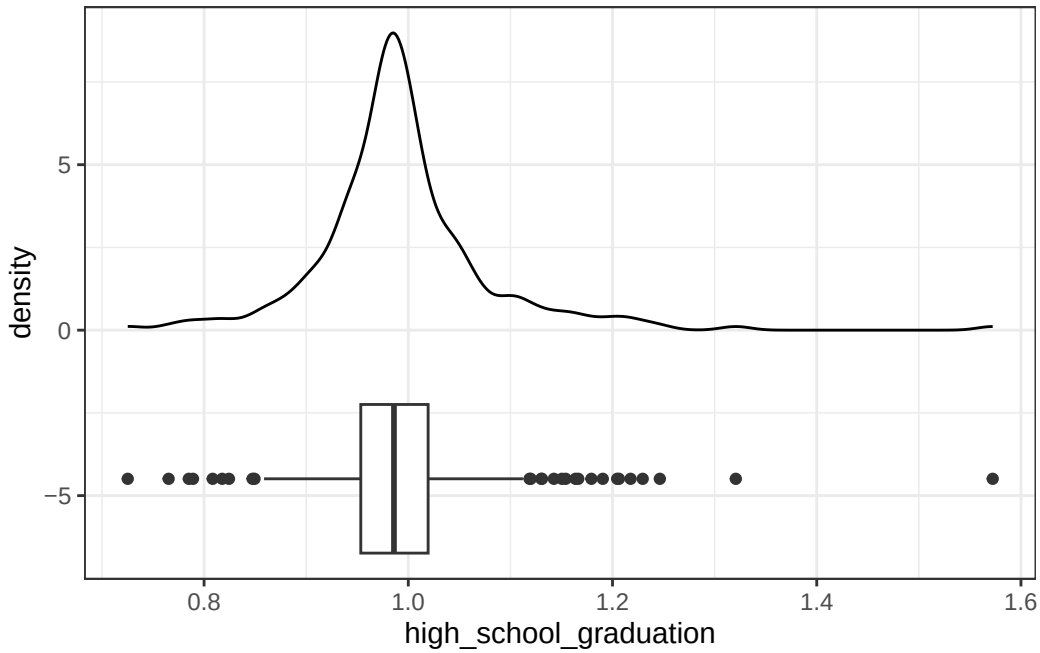

```
[[2]]
```

```
Warning: Removed 1 rows containing non-finite values (`stat_boxplot()`).
```

```
Warning: Removed 1 rows containing non-finite values (`stat_density()`).
```


density
traffic_volume
2
1
0
−1
0
5
10
15
20

```
[[3]]
```


density
high_school_graduation
5
0
−5
0.8
1.0
1.2
1.4
1.6

variable	mean	sd	min	p25	median	p75	iqr	mad	max	nr_missing
%_not_proficient_in_english	1.046	0.533	0.105	0.872	0.983	1.115	0.243	0.176	17.956	41
high_school_graduation	0.994	0.087	0.725	0.953	0.986	1.019	0.066	0.050	1.572	0
traffic_volume	1.069	0.730	0.000	0.844	0.966	1.109	0.265	0.192	20.129	1