

# Milestone 8: To Turn In

Maria Burzillo

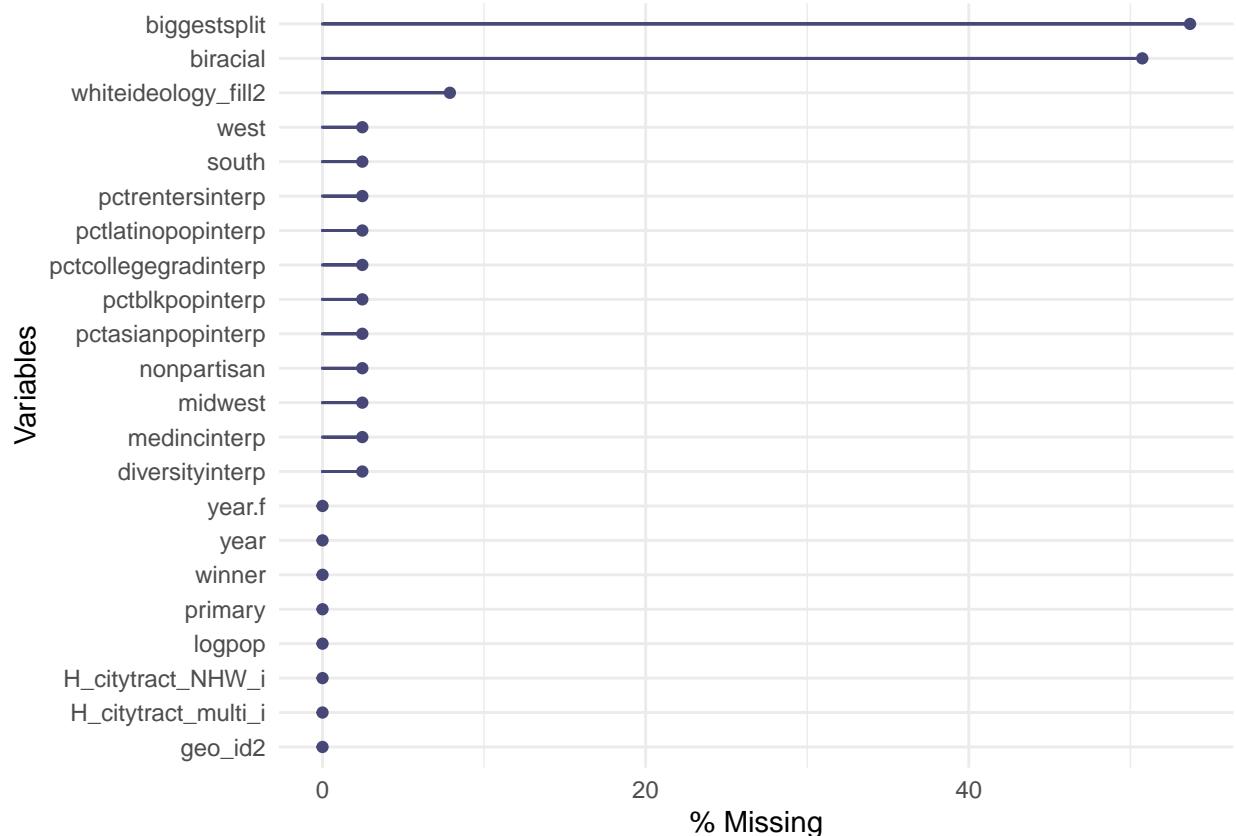
3/28/2020

## Abstract

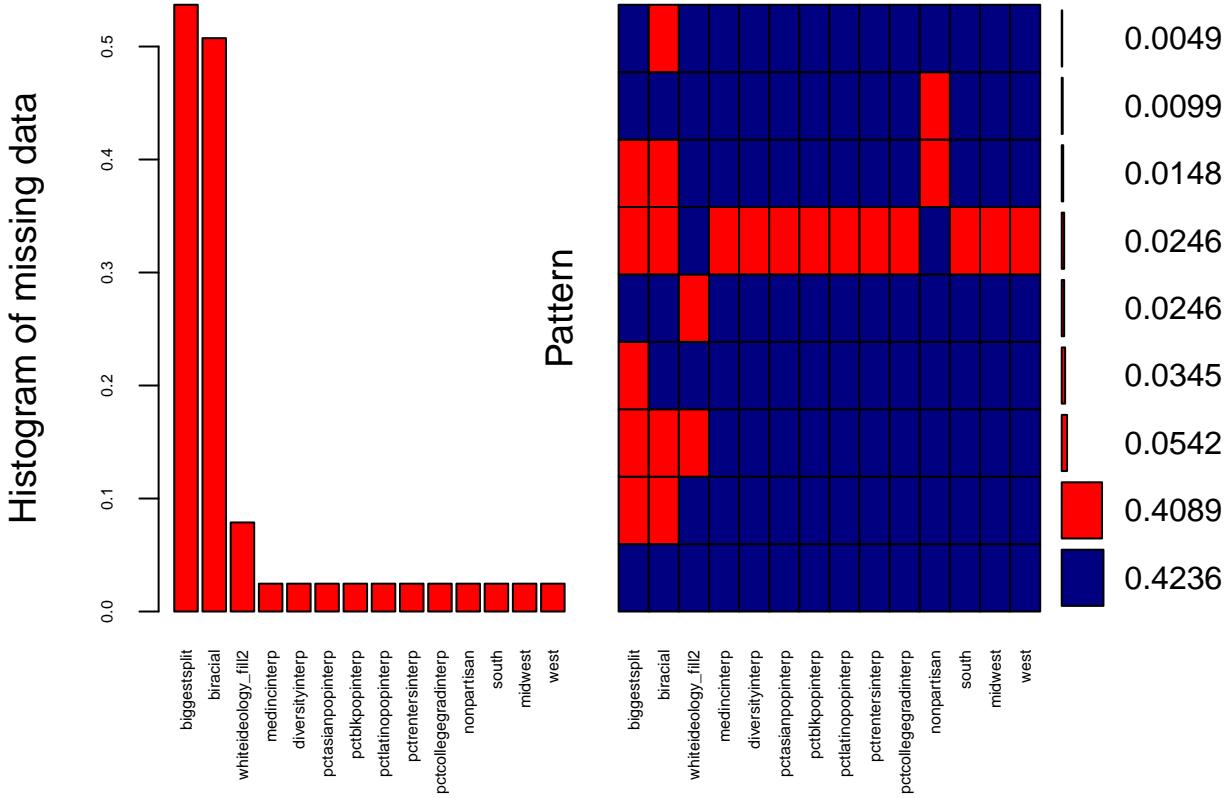
This is an extension of Jessica Trounstine's "Segregation and Inequality in Public Goods" (2016). I was able to replicate the main results of Trounstine's paper in R to suggest that racial segregation contributes to political polarization and decreased spending on public goods. Additionally, I extend the analysis by imputing missing data and rerunning Trounstine's original model as a robustness check.

## Extension 1

One aspect of Trounstine's paper with room for improvement is that there is a large amount of missing data in her datasets upon which she bases her analyses. Using R's mice package (add citation), we can perform multiple imputation. By performing imputation multiple times, this helps account for the uncertainty inherent in the individual imputations. Before performing the multiple imputations, we will first look at the missing data to see if there are any patterns.



The above plot reveals that `biggestsplit` and `biracial` are the two variables with by far the largest percentage of missing values. Both of these are missing approximately 50% of their observations. Since `biggestsplit` is the dependent variable in our analysis with this dataset, this is an important fact.



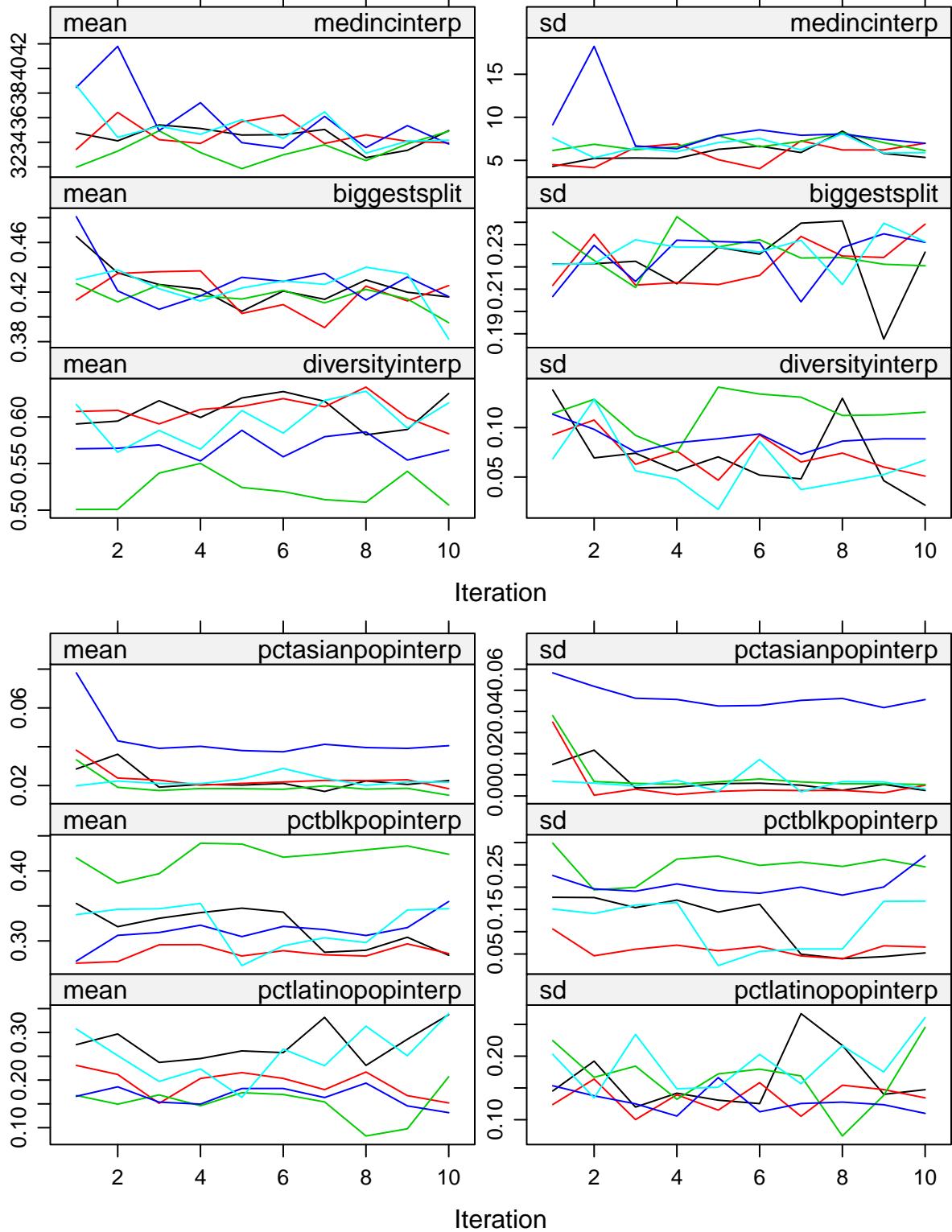
Variables sorted by number of missings:

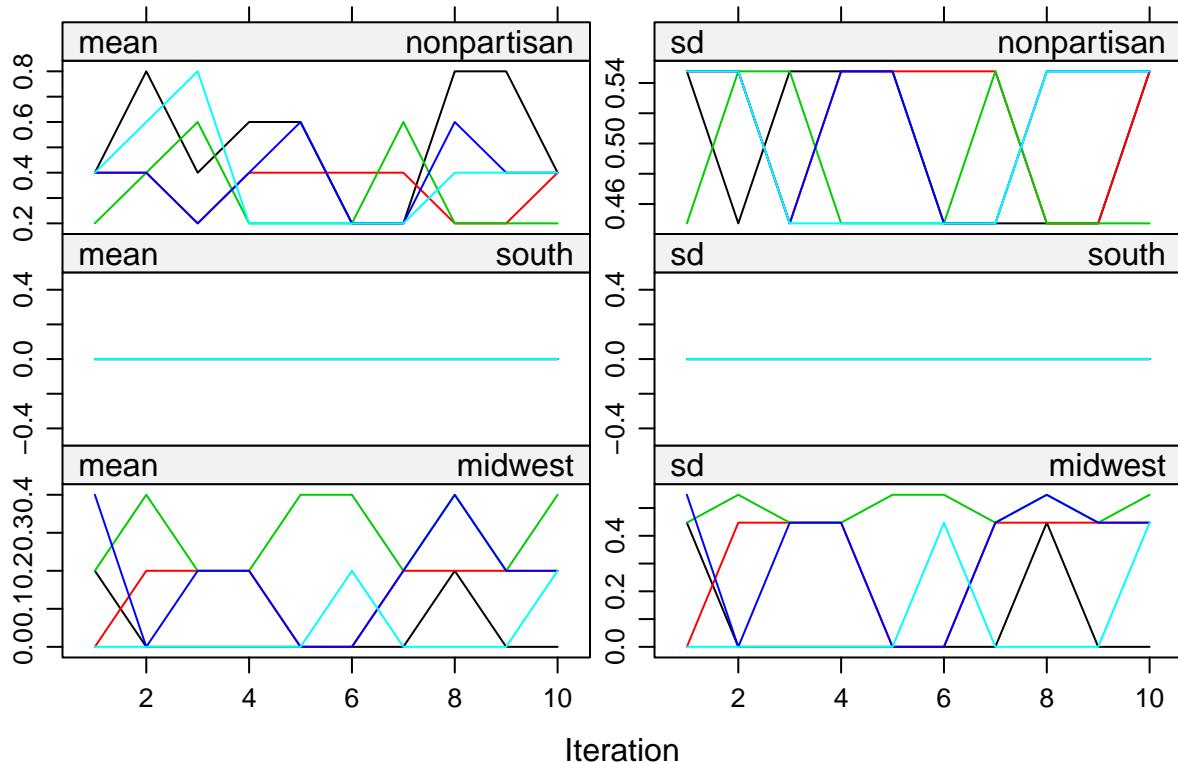
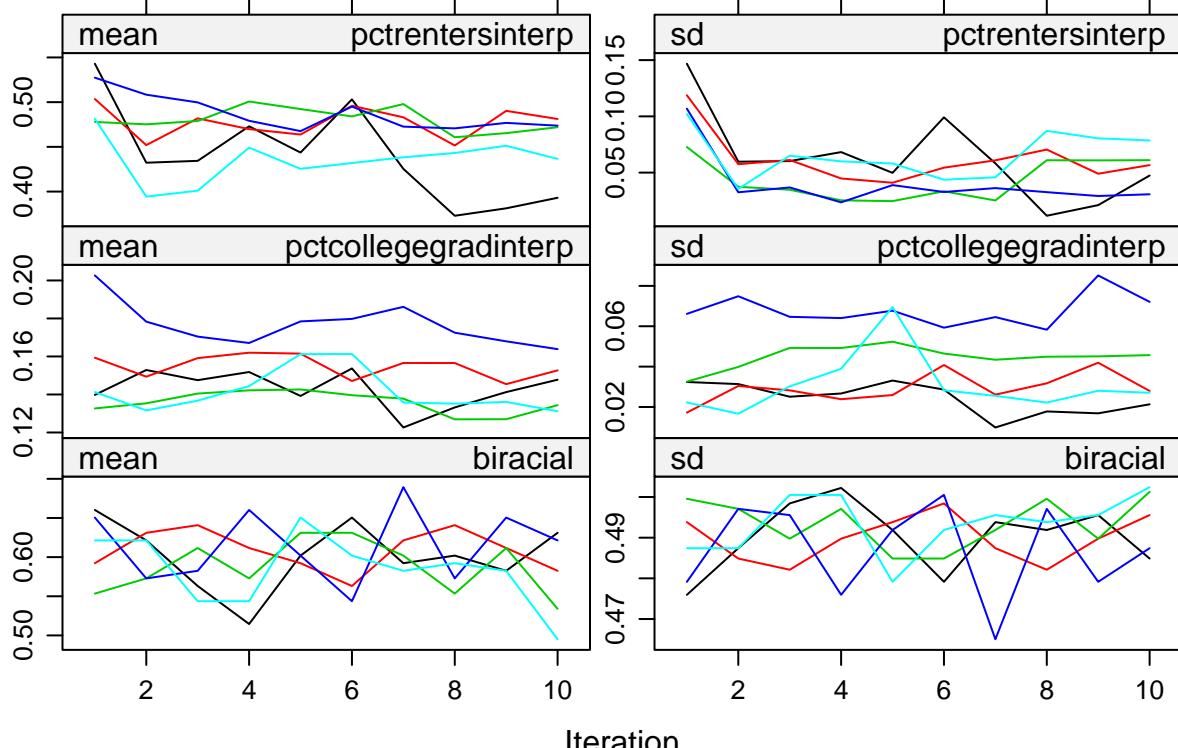
Variable	Count
biggestsplit	0.53694581
biracial	0.50738916
whiteideology_fill2	0.07881773
medincinterp	0.02463054
diversityinterp	0.02463054
pctasianpopinterp	0.02463054
pctblkpopinterp	0.02463054
pctlatinopopinterp	0.02463054
pctrentersinterp	0.02463054
pctcollegegradinterp	0.02463054
nonpartisan	0.02463054
south	0.02463054
midwest	0.02463054
west	0.02463054

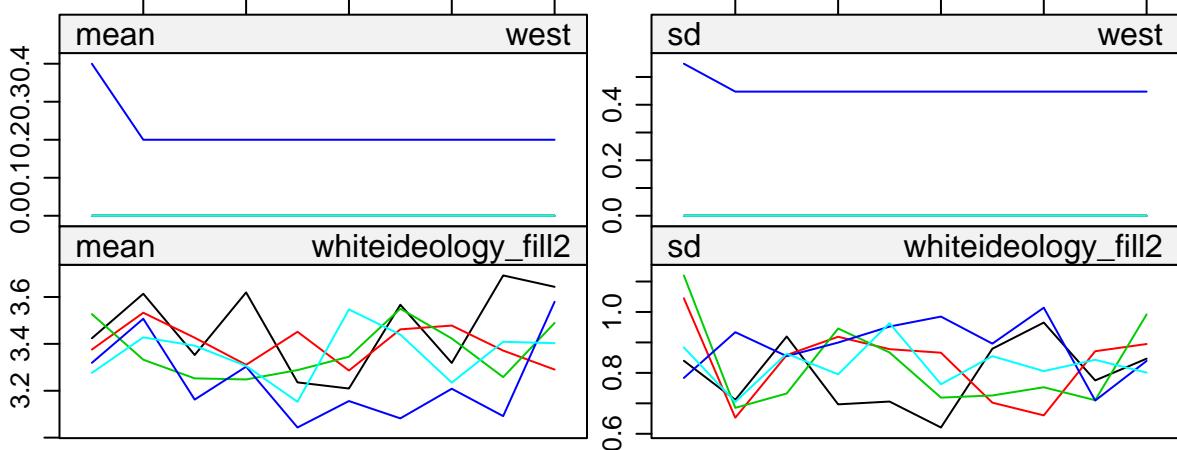
To better understand any potential patterns in missing data, we then plotted the pattern of missingness for only those variables missing values. From the plot on the right, we can see that approximately 42% of observations are complete. There seem to be a correspondence between missing a value for biggest split and missing biracial. There also seem to be about 2% of values for which most of the variables are missing. However, most observations are not missing more than 2-3 values.

Finding no clear patterns in the missing data, I next performed multiple imputations (with 10 iterations) on the dataset. A non-stochastic imputation method, Classification and Regression Trees (CART), was used instead of the default because of an error with matrix inversion caused by the data. Before examining the results of Trounstein's model using the imputed data, I first run some diagnostic tests of the imputation results to make sure that everything is running as expected.

First, I check the convergence of the algorithm used within mice() for each of the variables. For the most part, the fits intertwine and do not exhibit any trends at later iterations, as desired. There are some issues with the results for some of the variables with very few missing values (such as West and South), which makes sense as we would expect the mean to be less reliable due to the law of large numbers, and convergence is more difficult. Nevertheless, since there are so few of these values in the actual dataset, this is not a major concern.

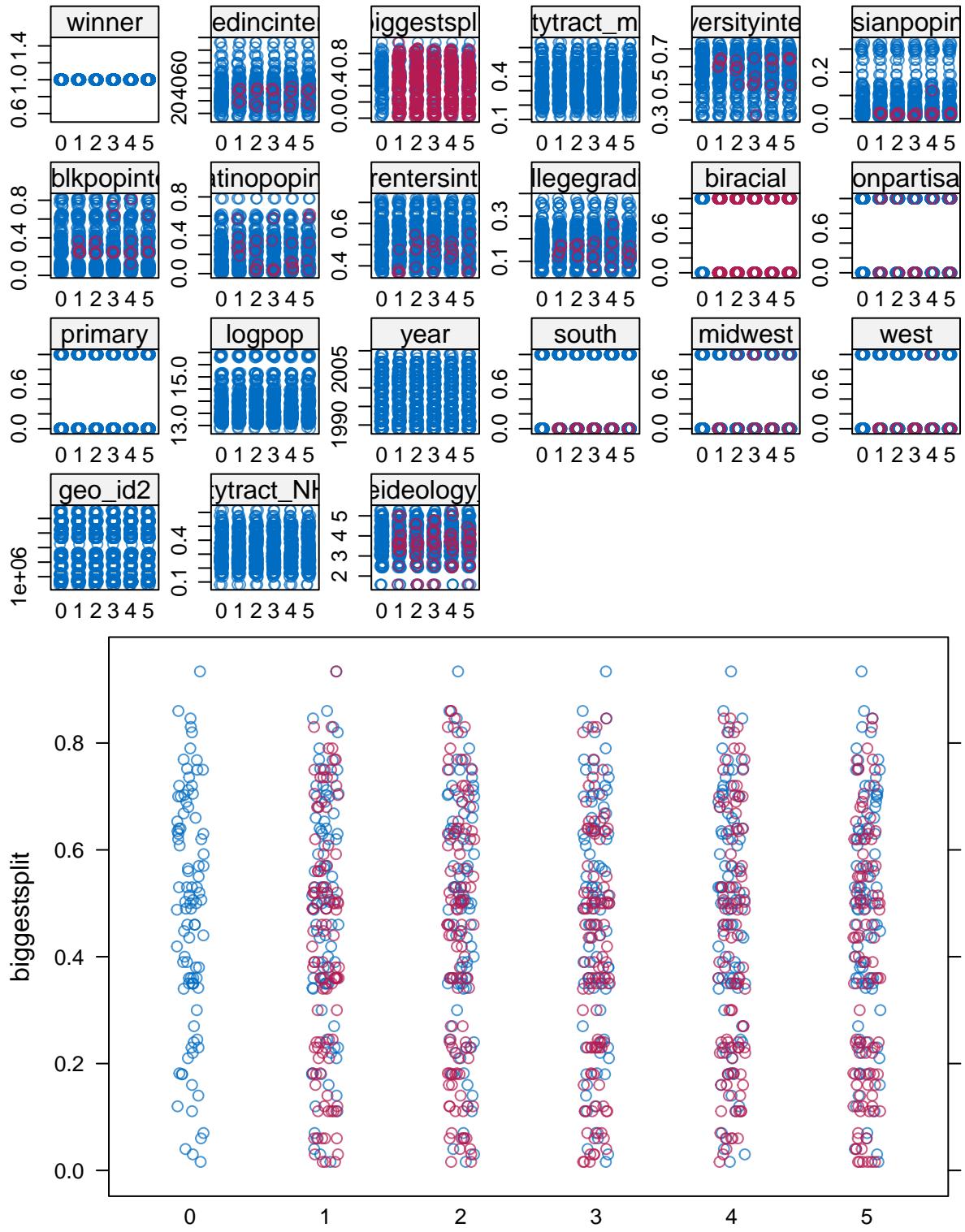




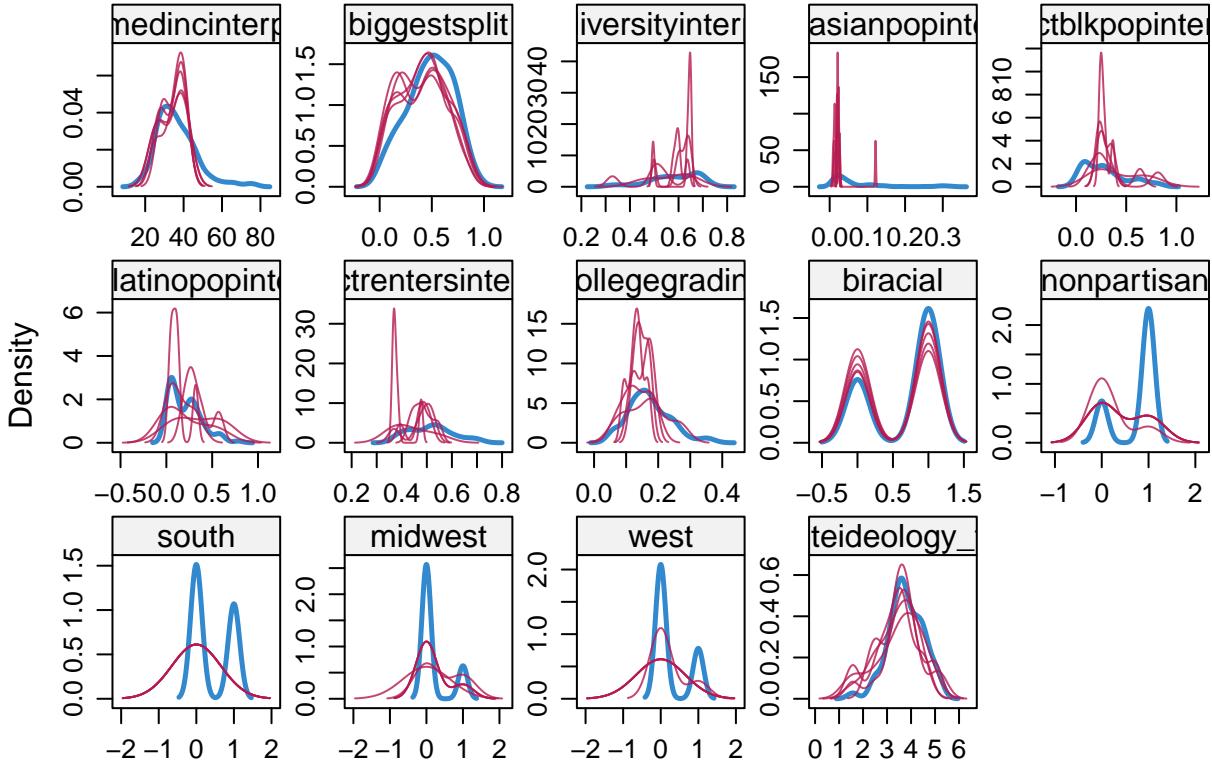


### Iteration

We can also check the imputed values against the original values using `stripplot()`. Each column in each subplot represents a separate iteration. The magenta points represent the imputed data. The values of the variable in question are along the y axis. We expect the spread of the data to be similar if the imputations were done well. If the data were missing completely at random, then the imputed data should have the same distribution as the original data. In particular, we want to be sure that the imputations are within a plausible range of the data. This is the case for all of our imputed variables, and there does not seem to be cause for alarm from these results.



Finally, we can also look at the density plots for each variable's actual data and for their imputed data from each of the iterations, which are represented in magenta. Overall, the density plots align quite well for the variables with the most missing data, `biracial` and `biggestsplit`, and relatively well for most of the other variables with missing data. Again, the fit is less good for variables with fewer missing data points.



While the imputations are not perfect, there do not seem to be any major problems so far. Thus, we can now proceed with our analysis. We fit each of our ten imputed datasets to Trounstein's 3 models using this data and then pool the results for each. The results are as follows:

## Extension 1 Racial Polarization Table

### Regression 1

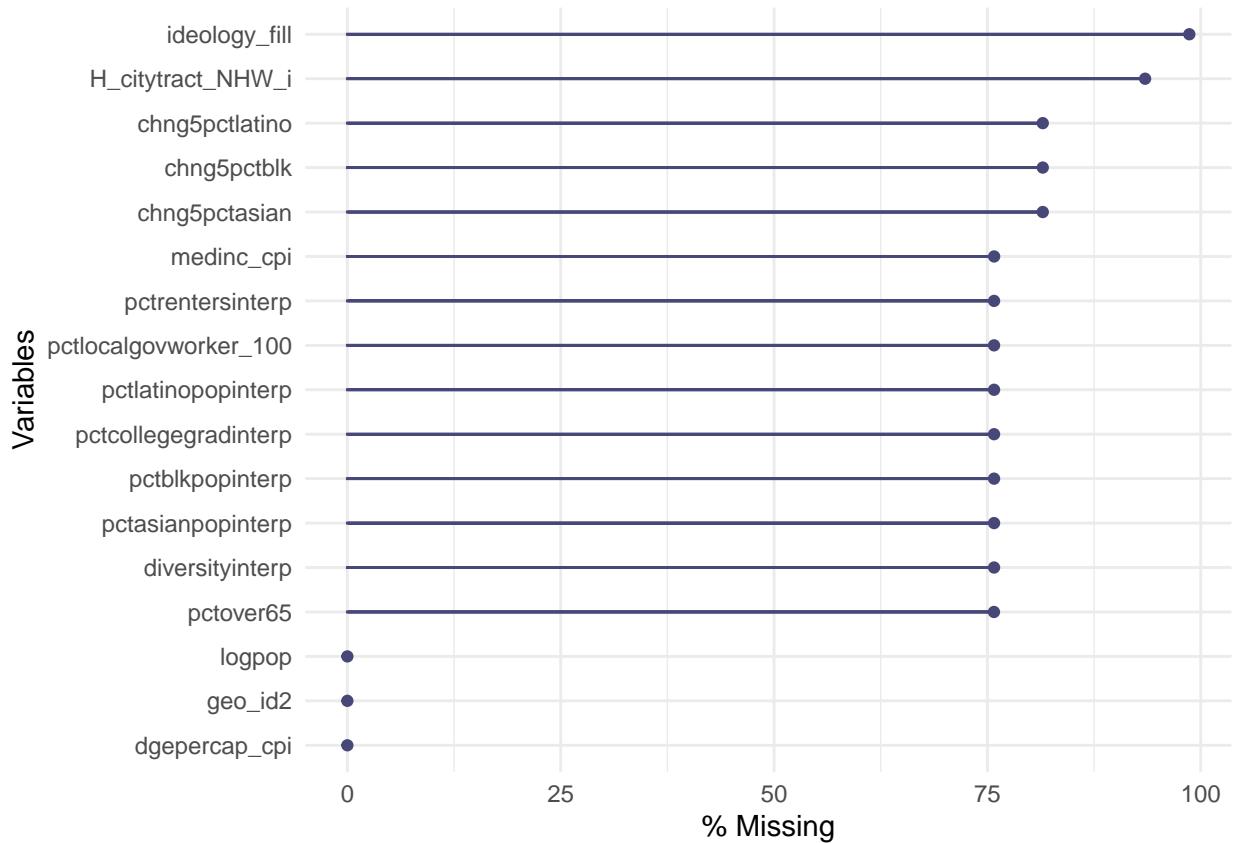
	term	estimate	std.error	p.value
1	(Intercept)	-0.017551361	0.443446057	0.9688597432
2	H_citytract_multi_i	0.438475039	0.364179593	0.2484118568
3	diversityinterp	0.261742932	0.330131866	0.4380653505
4	pctasianpopinterp	-0.115678535	0.466462999	0.8077637516
5	pctblkpopinterp	0.195651828	0.225407514	0.4033840742
6	pctlatinopopinterp	0.142738080	0.168579939	0.4012975487
7	medincinterp	-0.001112331	0.003941342	0.7790298356
8	pctrentersinterp	-0.359067684	0.258655257	0.1696400961
9	pctcollegegradinterp	0.230834704	0.518846662	0.6608054801
10	biracial	0.200931551	0.038427077	0.0003854007
11	nonpartisan	-0.075110241	0.060864522	0.2378265369
12	primary	-0.070496970	0.024177816	0.0045302323
13	logpop	0.013832569	0.033635438	0.6841505664

```
% Error: Argument 'header' must be of type 'logical' (TRUE/FALSE)
```

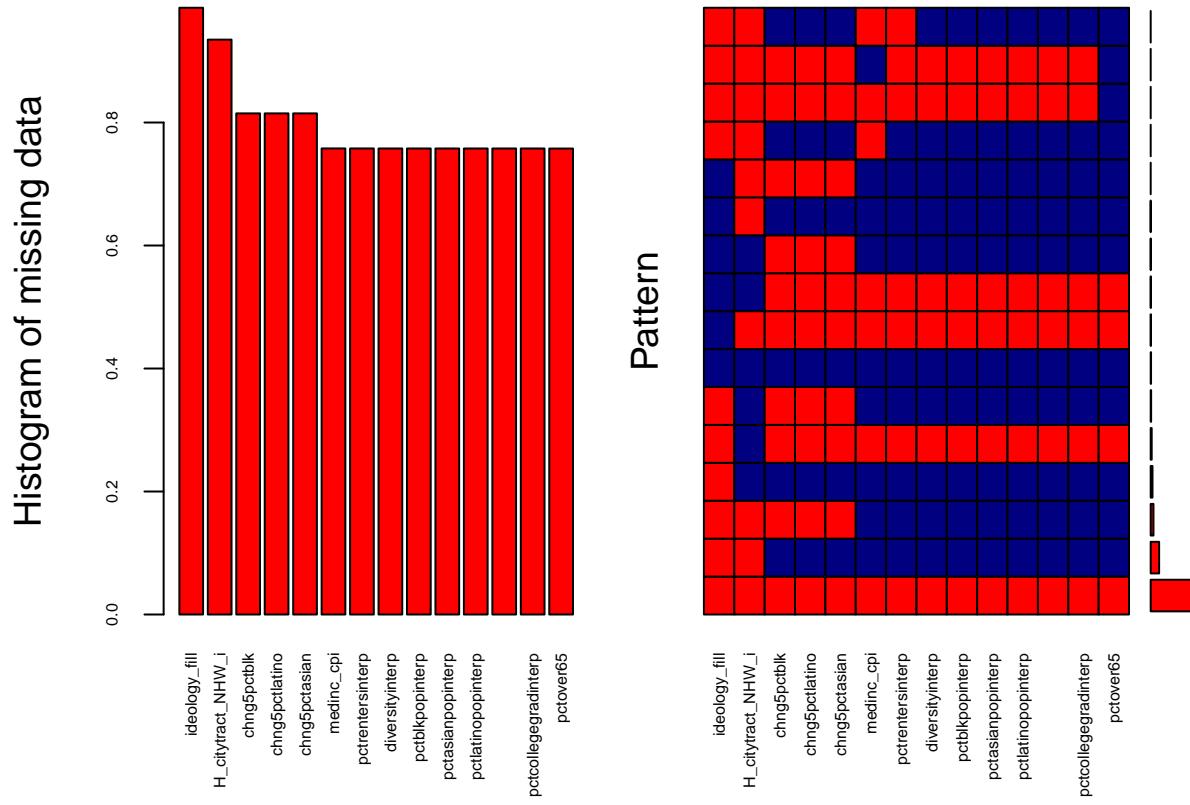
The results of the model with the new dataset are slightly different in comparison to the original results from Trounstein (2016). With the imputed data, we now have a total of 203 observations in our model as compared to the original model, which had only 91 observations. Interestingly, while the sign of the coefficient on the



there are a whopping 282,334 missing values, or %87.09 of the data.



Warning in plot.aggr(res, ...): not enough horizontal space to display frequencies



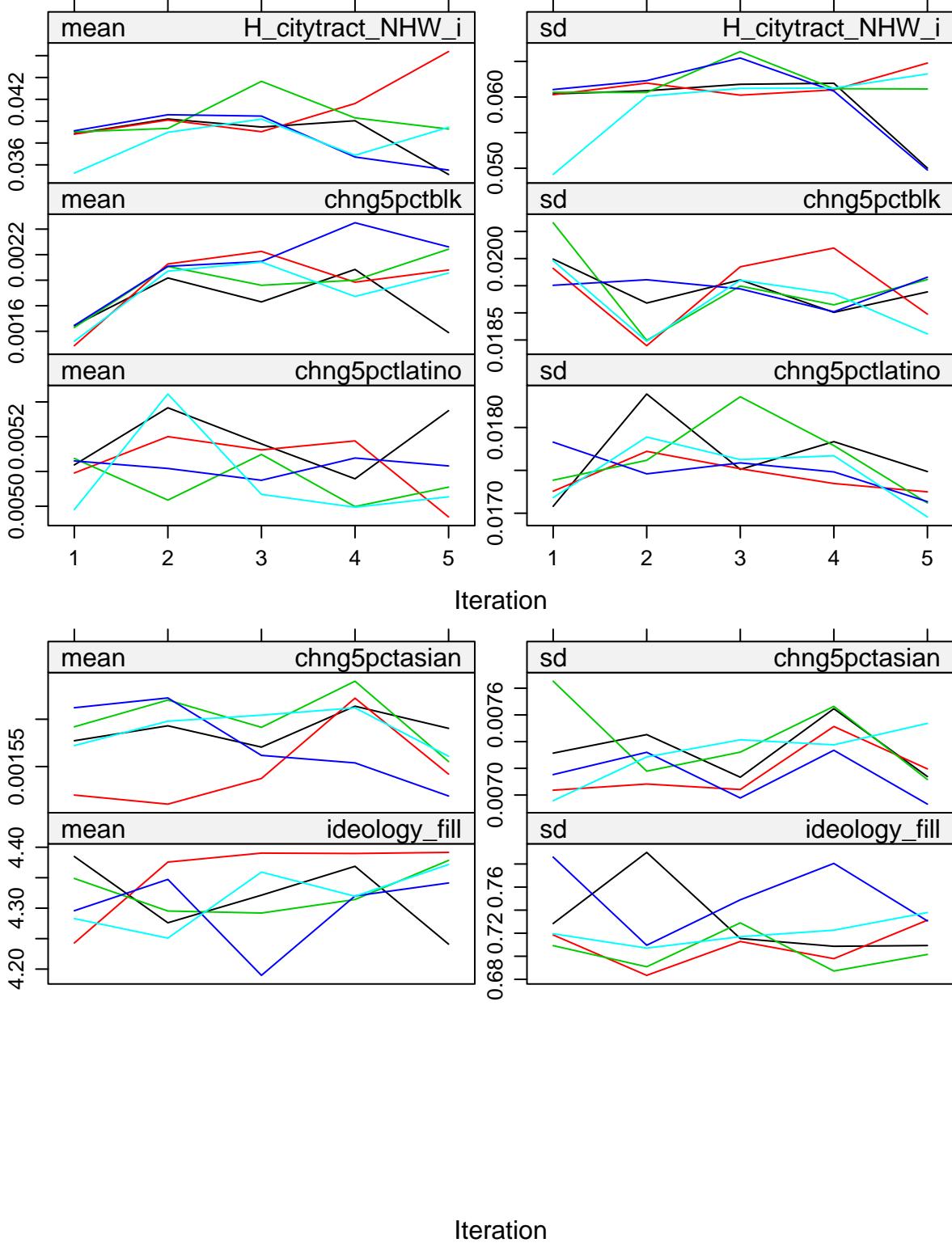
Variables sorted by number of missings:

Variable	Count
ideology_fill	0.9866041
H_citytract_NHW_i	0.9347787
chng5pctblk	0.8148547
chng5pctlatino	0.8148547
chng5pctasian	0.8148547
medinc_cpi	0.7578741
pctrentersinterp	0.7577980
diversityinterp	0.7577914
pctblkpopinterp	0.7577914
pctasianpopinterp	0.7577914
pctlatinopopinterp	0.7577914
pctlocalgovworker_100	0.7577914
pctcollegegradinterp	0.7577914
pctover65	0.7576722

[1] 1751.791

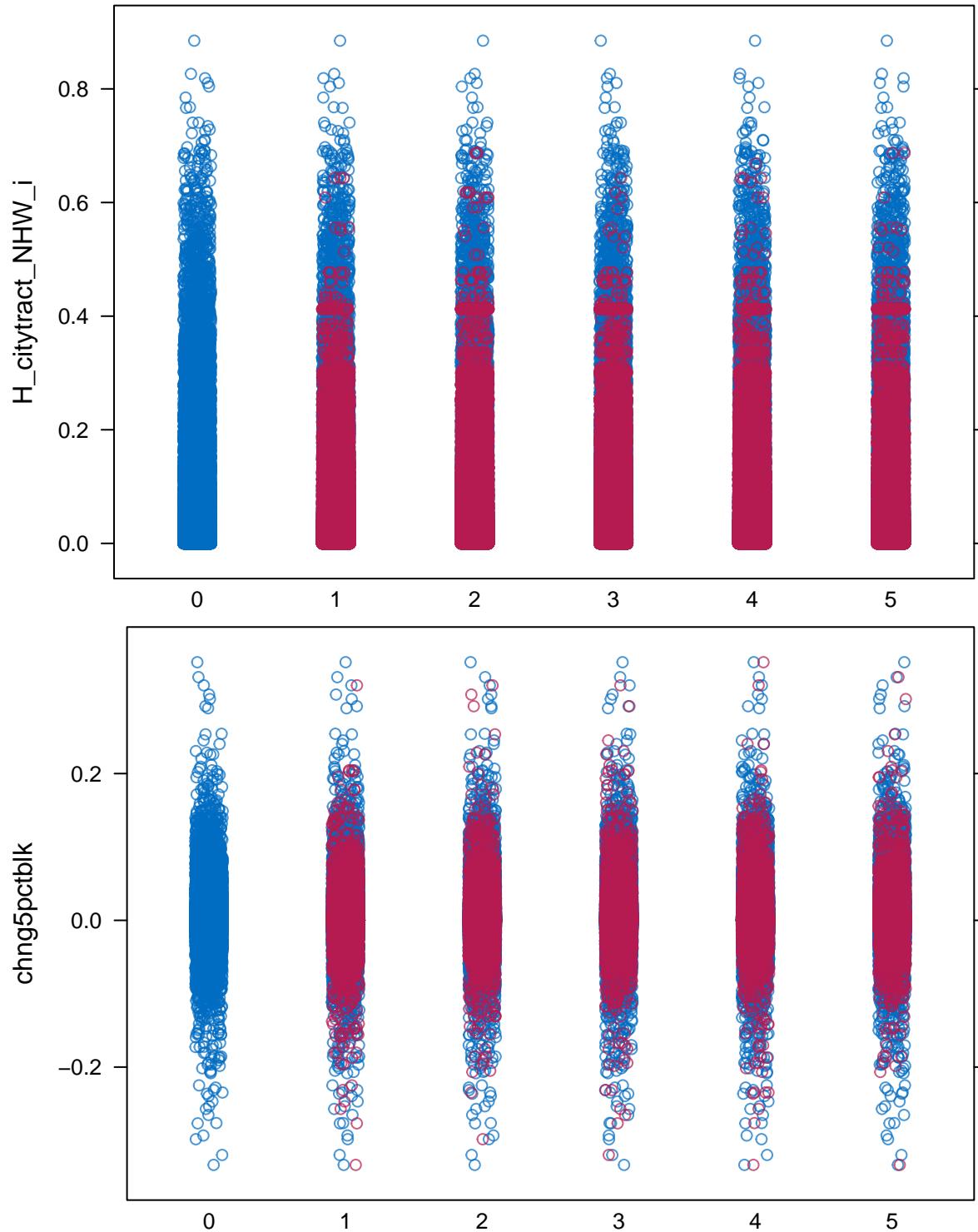
Finding no clear patterns in the missing data, I next performed multiple imputations (with 5 iterations) on the dataset. A non-stochastic imputation method, Classification and Regression Trees (CART), was used instead of the default because of an error with matrix inversion caused by the data. Before examining the results of Trounstein's model using the imputed data, I first run some diagnostic tests of the imputation results to make sure that everything is running as expected.

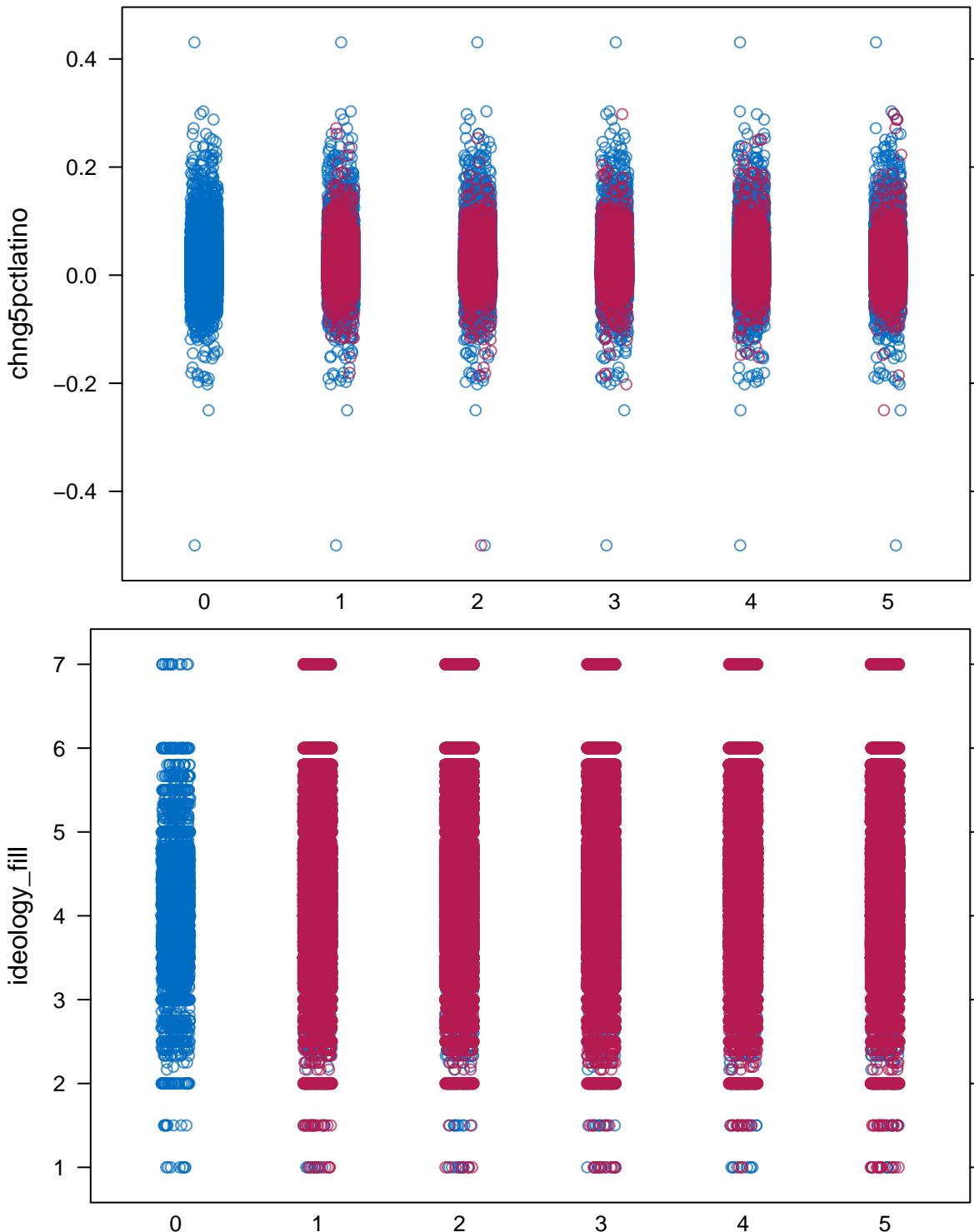
First, I check the convergence of the algorithm used within mice() for each of the variables. For the most part, the fits intertwine and do not exhibit any trends at later iterations, as desired.



We can also check the imputed values against the original values using `stripplot()`. Each column in each subplot represents a separate iteration. The magenta points represent the imputed data. The values of the variable in question are along the y axis. We expect the spread of the data to be similar if the imputations were done well. If the data were missing completely at random, then the imputed data should have the same distribution as the original data. In particular, we want to be sure that the imputations are within a plausible

range of the data. This is the case for all of our imputed variables, and there does not seem to be cause for alarm from these results.





Finally, we can also look at the density plots for each variable's actual data and for their imputed data from each of the iterations, which are represented in magenta. Overall, the density plots align quite well for the variables.

```
Error in density.default(x = c(NA_real_, NA_real_, NA_real_, NA_real_, : need at least 2 points to sele
```



```

6      chng5pctlatino -1.014508311 0.793333284 2.028593e-01
7      chng5pctasian  1.686429648 2.517734486 5.040126e-01
8      medinc_cpi    0.005921039 0.006684572 3.757500e-01
9  pctlocalgovworker_100 0.030462485 0.035456494 3.902690e-01
10     pctrentersinterp 0.144557921 0.331668113 6.629502e-01
11     pctover65       0.534087883 0.774798577 4.906281e-01
12  pctcollegegradinterp 3.411800021 0.818651976 3.093304e-05
13      logpop        -0.584239699 0.146492324 6.684994e-05

              term      estimate    std.error    p.value
1  H_citytract_NHW_i  0.057633818 0.358646668 8.753131e-01
2  diversityinterp   0.757071694 0.294414820 1.013574e-02
3  pctblkpopinterp   0.484826286 0.350278417 1.663405e-01
4  pctasianpopinterp 0.964382001 1.262549226 4.449746e-01
5  pctlatinopopinterp 1.573991769 0.441006106 3.591862e-04
6  medinc_cpi        0.005685258 0.006677415 3.945500e-01
7  pctlocalgovworker_100 0.031130824 0.035699427 3.832063e-01
8  pctrentersinterp   0.080869558 0.333815976 8.085832e-01
9  pctover65          0.609496623 0.775584389 4.319633e-01
10  pctcollegegradinterp 3.128261863 0.784705963 6.731650e-05
11      logpop        -0.588712905 0.146981960 6.218445e-05
12  ideology_fill    -0.014399020 0.013560657 2.901649e-01

```