Missing_Values_Treatment

Marta Fajlhauer

17 April 2018

Missing data imputation methods theory

The very first step in data analytics process is to clean variables from missing values by imputation whenever there is a reasonable possibility to do so, for example, if there is less than 50 percent of data missing or removing the variable from the dataset before final modelling process if the missingness is untreatable. The raw dataset has many missing values. This is important to recognise method that we should use for missing values treatment as it informs us about the options that we should use to impute or remove the missing data. There are three main categories of missing data mechanisms:

- **Missing Completely At Random (MCAR)**, occurs when data's missingness is unrelated to the data.
- **Missing At Random(MAR)**, means there is a systematic relationship between the missingness of an outcome variable' and other observed variables, but not the outcome variable itself.
- **Missing Not At Random (MNAR)**, occurs when a variable's missingness is related to the variable itself. It is the weight of the observation itself that is the cause of its being missing. It's like in for some variables in our datasets for which we have the full record but after removing observations recorded as 0 or 1 we have obtained missing values. From the information given on the website from where we have obtained the dataset, we know that: *all values more than 3 SD above the mean are normalized to 1.00; all values more than 3 SD below the mean are normalized to 0.00*[1] so these observations weight is the cause to be removed.

There are different methods of data imputation.

Mean substitution:

The benefit of this method is that it produces unbiased estimates of the mean of a column, however, it produces biased estimates of the variance, since it removes the natural variability that would have occurred in the missing values if they would be not missing. The equation for the substitution is

$$\bar{x} = \frac{\sum x}{n}$$

Regression imputation

The benefit of this method is that it produces unbiased estimates of the mean and regression coefficients (so long as the relevant variables are included in the regression model). This improves variability in the data as the predicted values of the missing data lie right on the regression line but, as we know, very few data points lie right on the regression line—there is usually a normally distributed residual (error) term. Due to this, regression imputation underestimates the variability of the missing values. Hence it will result in biased estimates of the variance and covariance between different columns. Since we expected the error term to be zero we obtain the equation:

$$\widehat{y} = X\beta = X(X^T X)^{-1} X^T y$$

Stochastic regression method

The benefit of this method is that it produces unbiased estimates of the mean, variance, covariance, and regression coefficients. It does this by adding a random (stochastic) value to the predictions of regression imputation. This random added value is sampled from the residual (error) distribution of the linear regression—which is assumed to be a normal distribution. This method restores the variability in the missing values. Drawbacks are that standard errors and confidence intervals are smaller than they should be. Since it

produces only one imputed dataset, it does not capture the extent to which we are uncertain about the residuals and our coefficient estimates. Since the error term (epsilon) we do not expect to be zero it is now included in the calculation:

$$\widehat{y} = X\beta = X(X^T X)^{-1} X^T y + \epsilon$$

Multiple imputation

We generate multiple versions of the imputed data with different estimations of the missing data, where the imputed values are drawn from a distribution. The uncertainty about what the imputed values should be is reflected in the variation between the multiply imputed datasets. We perform our intended analysis separately with each of these m amounts of completed datasets.

Predictive mean matching (pmm) in R

The first step performs stochastic linear regression imputation using coefficients for each predictor estimated from the data. The second step chooses slightly different estimates of these regression coefficients and proceeds into the next imputation. The first step of the next imputation uses the slightly different coefficient estimates to perform stochastic linear regression imputation again. After that, in the second step of the second iteration, still, other coefficient estimates are generated to be used in the third imputation. This cycle goes on until we have m multiply imputed datasets. How do we choose these different coefficient estimates at the second step of each imputation? Traditionally, the approach is Bayesian in nature; these new coefficients are drawn from each of the coefficients' posterior distribution, which describes credible values of the estimate using the observed data and uninformative priors. [2]

Missing data imputation methods practice

First decision made was to remove county and community variables due to many missing values and the fact that there are variables community name and state that represent a similar thing. I also removed fold variable as I won't be concentrating on unfolding data. I constructed a table showing pattern of missing data.

```
     state communityname population householdsize PolicCars PolicOperBudg
 319     1             1          1             1         1             1
1675     1             1          1             1         0             0
         0             0          0             0      1675          1675
     LemasPctPolicOnPatr LemasGangUnitDeploy PolicBudgPerPop
 319                   1                   1               1     0
1675                   0                   0               0    22
                    1675                1675            1675 36850
```

In the table the rows are sorted in an increasing amount-of-missingness order, the first row always refers to the missing data pattern containing the least amount of missing data. From this table, we can see that USA Census data has a full record but the FBI and LEMAS data contain many missing values. In the table, the last row contains a count of the number of missing data points in each column. From the data collection description we also know that *all values more than 3 SD above of the mean are normalized to 1.00, all values more than 3 SD below the mean are normalized to 0.00*[1]. Hence they are all potential outliers so the decision to remove those values. After that, I constructed the missing data pattern table once again. Due to the size of the table: 546 rows and 126 columns. I have summarised number of missing values for each variable.

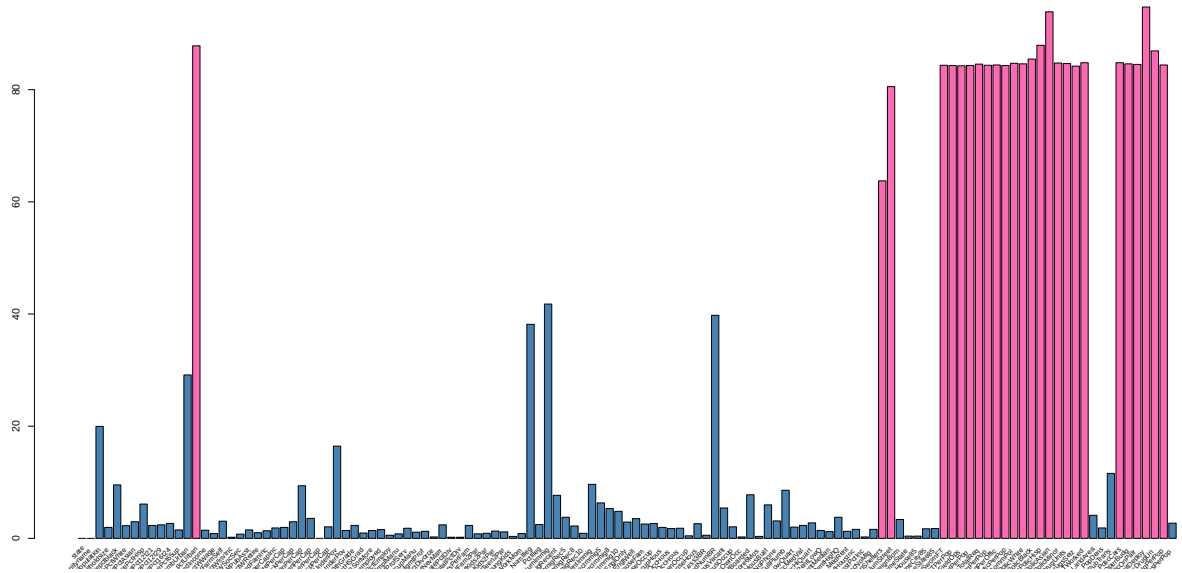
```
               state          communityname           OtherPerCap
                   0                      0                     0
          pctWInvInc           FemalePctDiv          TotalPctDiv
                   4                      4                     4
       MalePctDivorce          PctHousOwnOcc      MedOwnCostPctInc
                   5                      5                     5
  PctWorkMomYoungKids         PctVacMore6Mos      PctBornSameState
                   7                      7                     8
       PctSameHouse85        PctPersOwnOccup             PctEmploy
                   8                      9                    11
```

PctHousLess3BR	pctWSocSec	PctEmplManu
11	15	16
PctFam2Par	pctWWage	PctWorkMom
16	17	17
PctKids2Par	PctImmigRec10	PctNotHSGrad
18	18	19
pctWRetire	PctOccupManu	PctTeen2Par
20	22	23
RentMedian	PctOccupMgmtProf	MedRent
24	25	25
PctYoungKids2Par	medFamInc	PctPopUnderPov
26	27	28
PctBSorMore	RentLowQ	medIncome
28	28	29
agePct65up	pctWPubAsst	PctUnemployed
30	30	31
MedRentPctHousInc	MedOwnCostPctIncNoMtg	PctSameCity85
32	32	34
PersPerOwnOccHous	PctSameState85	PctEmplProfServ
35	35	36
PersPerRentOccHous	perCapInc	PopDens
36	37	37
householdsize	whitePerCap	PersPerOccupHous
39	39	39
OwnOccLowQuart	HispPerCap	PctHousOccup
40	41	41
PctImmigRec8	racePctWhite	agePct12t21
44	45	46
PctLess9thGrade	PersPerFam	OwnOccMedVal
46	46	46
agePct12t29	MalePctNevMarr	PctIlleg
48	48	49
PctLargHouseFam	PctPersDenseHous	agePct16t24
51	52	53
PctLargHouseOccup	ViolentCrimesPerPop	OwnOccHiQuart
53	54	55
PctSpeakEnglOnly	racePctAsian	blackPerCap
58	59	59
pctWFarmSelf	PctHousNoPhone	PctForeignBorn
61	62	67
PctNotSpeakEnglWell	AsianPerCap	PctImmigRec5
70	71	75
RentHighQ	LandArea	PctRecImmig10
75	82	96
PctRecImmig8	HousVacant	MedYrHousBuilt
106	108	119
racePctHisp	PctRecImmig5	PctImmigRecent
122	126	153
PctVacantBoarded	PctWOFullPlumb	indianPerCap
155	171	187
racepctblack	PctRecentImmig	PctUsePubTrans
190	192	231
NumUnderPov	population	numbUrban
328	398	581

NumIlleg	MedNumBR	NumImmig
761	793	833
NumInShelters	NumStreet	NumKindsDrugsSeiz
1271	1606	1679
LemasSwFTFieldOps	LemasSwFTPerPop	LemasSwFTFieldPerPop
1680	1681	1681
PolicPerPop	LemasSwornFT	LemasTotReqPerPop
1681	1682	1682
PolicReqPerOffic	PolicBudgPerPop	LemasPctPolicOnPatr
1683	1683	1685
LemasTotalReq	PctPolicWhite	PolicOperBudg
1686	1687	1687
OfficAssgnDrugUnits	RacialMatchCommPol	PctPolicMinor
1688	1689	1690
PolicAveOTWorked	PolicCars	PctPolicBlack
1691	1691	1704
LemasPctOfficDrugUn	pctUrban	PctPolicHisp
1733	1751	1753
PctPolicAsian	LemasGangUnitDeploy	
1872	1889	

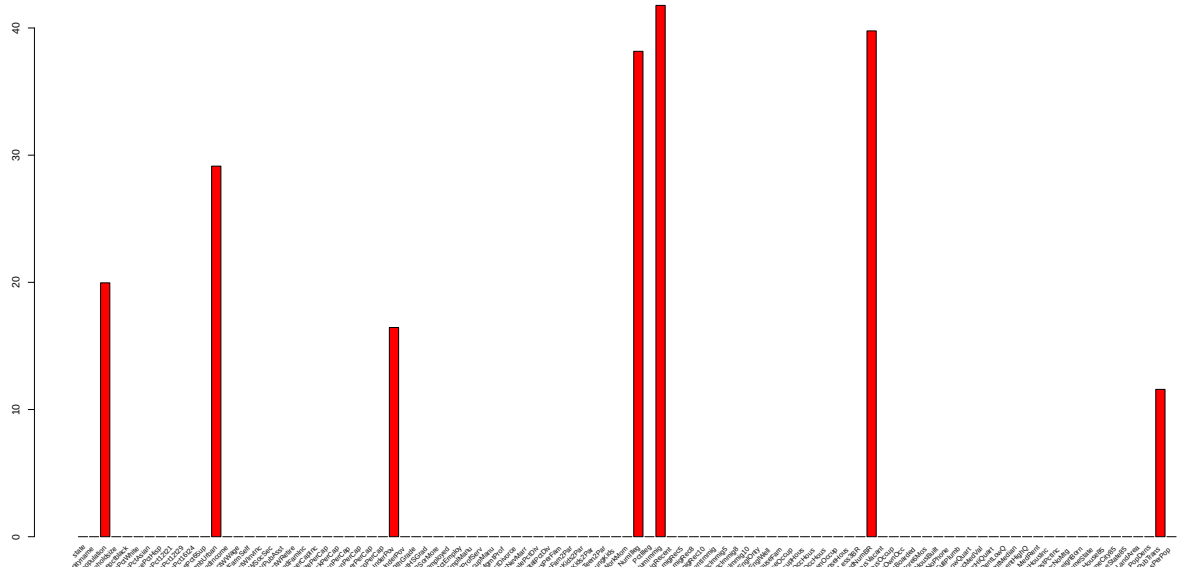
52203 missing numbers in the data, 21% of data consist of missing values. To simplify the selection process for data cleaning purposes we can visualise which variables we should choose. Variables that we marked as pink have more than 50 percent of missing values so we can remove them and recheck if we didn't omit any.

Figure 1 (a and b): missing data by percentage



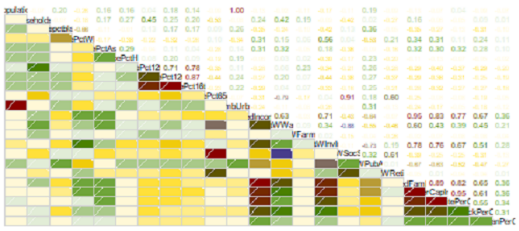
We can create a missing data pattern once again. Now there is less than 4% of missing values in the dataset. For those variables that missing values number is less than 10% we use mean substitution method. Whilst mean substitution produces unbiased estimates of the mean of a column, it produces biased estimates of the variance, since it removes the natural variability that would have occurred in the missing values had they not been missing. If we impute 10% of given variable by mean substitution method then we not make a big harm to our data but variables with more than 10% NA should be treated differently

Figure 1 (a and b): missing data by percentage

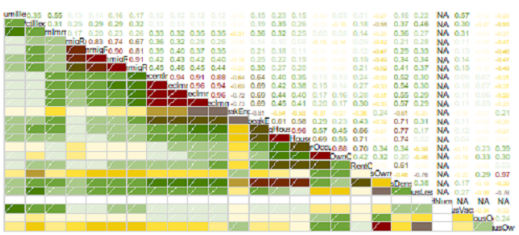


The variables: population, numbUrban, NumUnderPov, NumIlleg, NumImmig, MedNumBR, PctUsePubTrans have more than 10% of missing values hence we would like to use the predictive mean matching method in R. However first we will look into correlation matrix to remove variables that are highly correlated with each other. That's to remove colinearity in the data problem. Correlation matrices produced are:

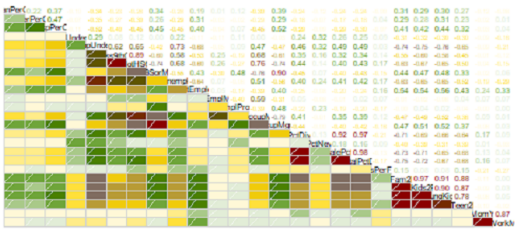
Vars 2 - Vars 25



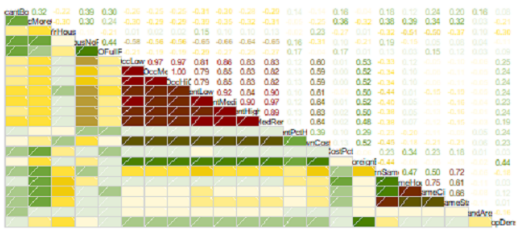
Vars 51 - Vars 75

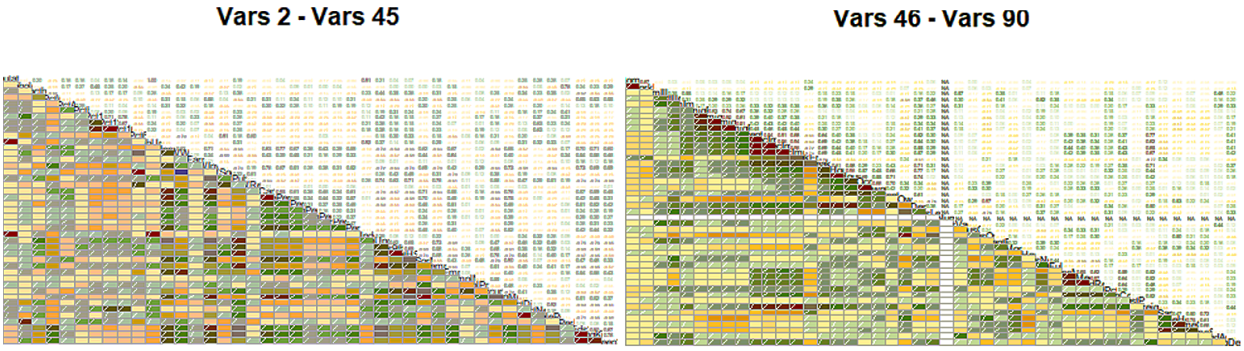


Vars 25 - Vars 50



Vars 76 - Vars 97





The full code to produce correlogram is:

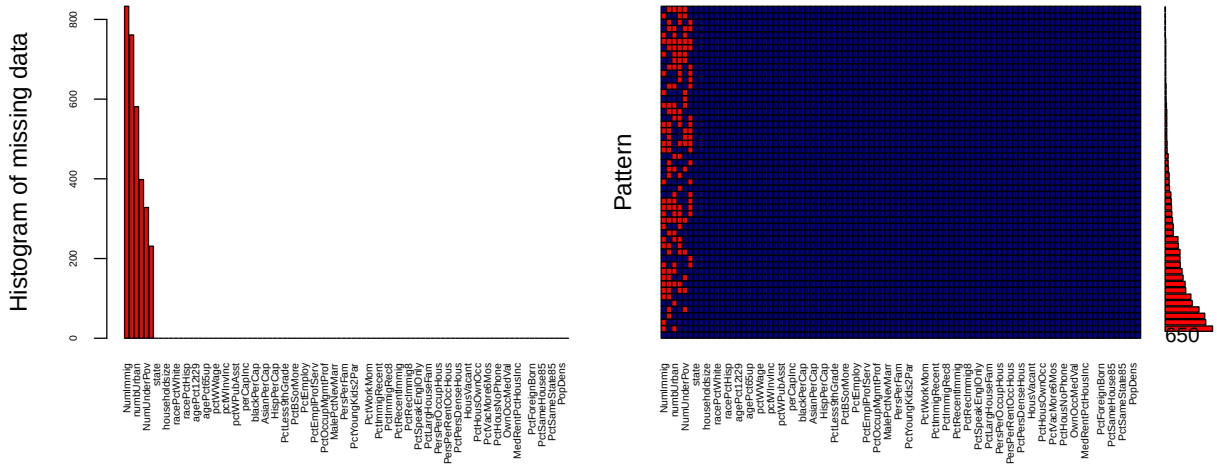
The decisions made based on the output of correlograms are:

- Median Family income is a subset of Median Income so decision to remove this variable.
- PctFam2Par and PctKids2Par are very similar things so Decision to remove PctFam2Par.
- Male PctDivorced and FemalePctDivorced are two partitions of Total percentage divorced hence decision to remove Total PctDivorced.
- OwnOccMedVla left but LowQ and HighQ removed.
- MedRent left but HighQ and LowQ removed

I have decided to construct correlogram once again after removing those variables to look for some future improvements.

- RentMedian (rental housing - median rent) and MedRent (median gross rent) measure the same feature so the decision of removing RentMedian.
- As medNumBR shows only 0.5 in non-missing values observations and there are 793 observations the variable is not taken into analysis

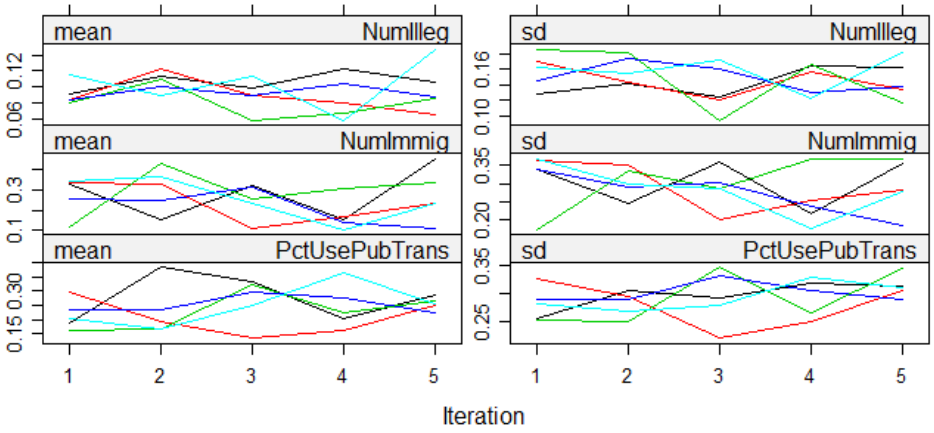
Figure 2: Visualisation of missing numbers



Once visualising the missing data we will apply the pmm method of imputation and visualise the output.

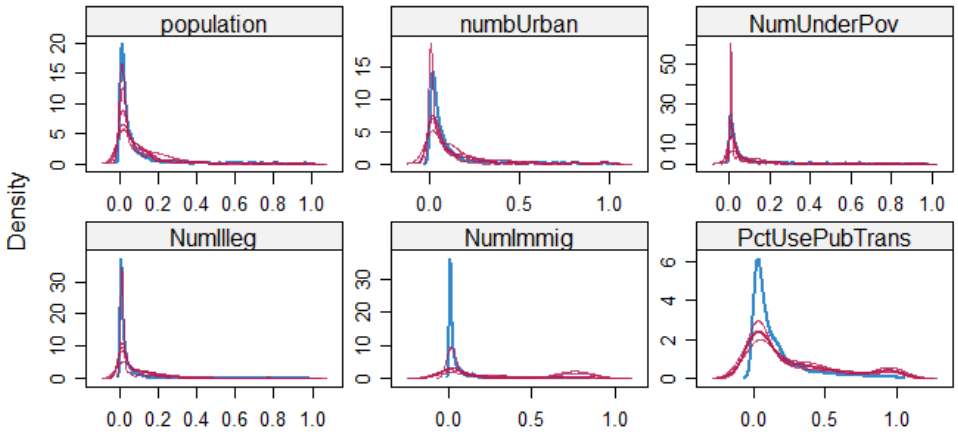
Multiple imputation method to impute data with missing values in multiple columns requires iteration over all these columns a few times. Every time R library called: "mice" produces imputations and using Bayes statistics sample new parameters estimates from the parameters posterior distribution for all columns with missing values data. The final imputation are the values from the final iteration. We can see trace plots of the mean and standard deviations for each variable with missing values where each line in each plot is one of the m imputations. We can't see the trend in the trace plots. Variance within a chain (m chains) should be equal to the variance between the chains that leads us to the conclusion that the convergence was achieved. If we would underlying trend in the plot for example that all plots on each graph space would merge together then the convergence would not be achieved.

Figure 3: Visualisation of the convergence in multiple imputation methods



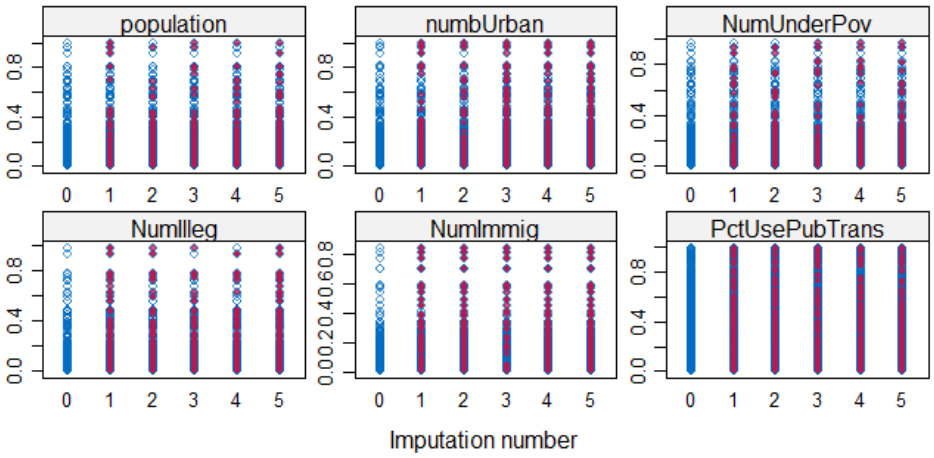
Next, we can plot density plot of the actual non-missing values (blue) and the imputed values (red plots). We can see that the distributions are similar

Figure 4: density plots of actual non-missing and the imputed values

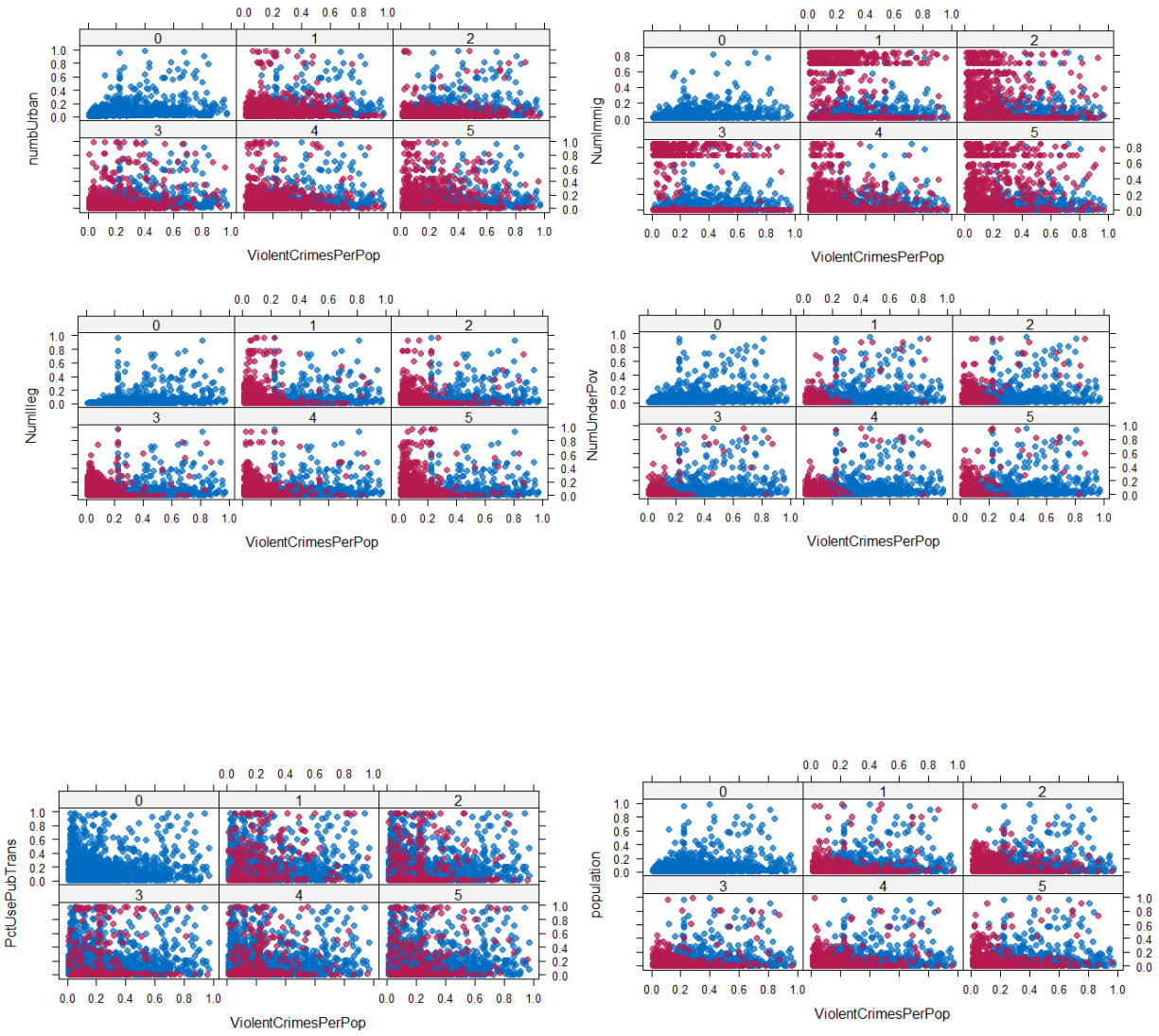


In addition to that, we can plot stripplot that shows the distributions of all imputed variables as individual points. The first plot shows missing values data and dataset produced after each imputation. Blue points are observed, the red points are imputed. Note that the red points follow the blue points reasonably well, especially in later iterations, including the gaps in the distribution.

Figure 5: stripplots of distribution of all imputed values



We can check how imputation process has inputted missing values at each variable during the iteration process.



Dataset selection process was made by comparing OLS result from 5 imputed dataset. Even if the dataset is not clean yet that was natural selection process since we want to perform advanced regression techniques after cleaning the dataset. With the highest R-squared, R-squared Adj and the smallest residual error dataset 1 seems to be the best choice hence decision of selecting this dataset for further analyses.

Figure 6: Table withimportant information to select imputed dataset

Dataset	R-squared	Adj R-sq	F-statistics	Residual
1	0.572	0.5524	29.28	0.1353
2	0.5713	0.5517	29.19	0.1355
3	0.5694	0.5497	28.97	0.1358
4	0.5704	0.5507	29.08	0.1356
5	0.5695	0.5499	28.98	0.1357

Reference

[1] UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/communities+and+crime

[2] R Predictive analysis Tony Fishetti, Eric Mayor, Rui Miguel Page 281 Packt books