Missing Value Imputation

Liu Wenhui Singh Ravi

3/27/2018

A. Missing Value Detection

A.1 Data preparation

```
suppressPackageStartupMessages({
library(Hmisc)
library(dplyr)
library(mice)
library(tidyr)
library(ggplot2)
library(purrr)
library(VIM)
library(magrittr)
library(corrplot)
})
```

A.1.1 Data Load

The dataset we used here is the air quality record of Beijing and in the dataset, there are different air pollution indicators such as PM2.5. For instance, there are 4 PM2.5 attributes (PM_Dongsihuan, PM_Nongzhguan, PM_Dongsi, PM_US.Post) which are PM 2.5 numbers collected from 4 different locations in Beijing.

```
air <- read.csv('BeijingPM20100101_20151231.csv')
air %<>% rename(PM_Dongsih= PM_Dongsihuan)%<>% rename(PM_Nongzh=PM_Nongzhanguan)
summary(air)
```

```
##        No             year          month             day
##  Min.   :    1   Min.   :2010   Min.   : 1.000   Min.   : 1.00
##  1st Qu.:13147   1st Qu.:2011   1st Qu.: 4.000   1st Qu.: 8.00
##  Median :26292   Median :2012   Median : 7.000   Median :16.00
##  Mean   :26292   Mean   :2012   Mean   : 6.524   Mean   :15.73
##  3rd Qu.:39438   3rd Qu.:2014   3rd Qu.:10.000   3rd Qu.:23.00
##  Max.   :52584   Max.   :2015   Max.   :12.000   Max.   :31.00
##
##       hour           season        PM_Dongsi        PM_Dongsih
##  Min.   : 0.00   Min.   :1.000   Min.   :  3.00   Min.   :  3.00
##  1st Qu.: 5.75   1st Qu.:1.000   1st Qu.: 24.00   1st Qu.: 28.00
##  Median :11.50   Median :2.000   Median : 64.00   Median : 68.00
##  Mean   :11.50   Mean   :2.491   Mean   : 89.15   Mean   : 92.56
##  3rd Qu.:17.25   3rd Qu.:3.000   3rd Qu.:124.00   3rd Qu.:127.00
##  Max.   :23.00   Max.   :4.000   Max.   :737.00   Max.   :672.00
##                                  NA's   :27532    NA's   :32076
##    PM_Nongzh        PM_US.Post         DEWP              HUMI
##  Min.   :  3.00   Min.   :  1.0   Min.   :-40.000   Min.   :  2.0
```

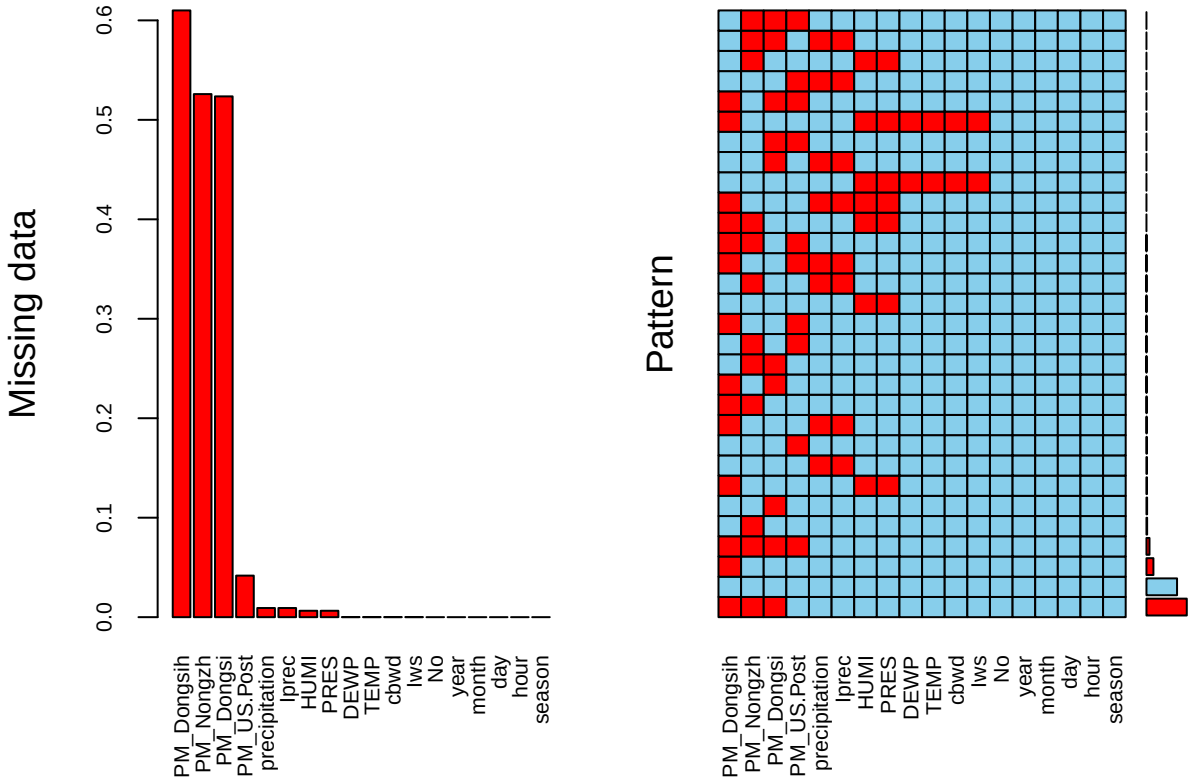
```
##  1st Qu.: 24.00   1st Qu.: 27.0   1st Qu.:-10.000   1st Qu.: 31.0
##  Median : 62.00   Median : 69.0   Median :  2.000   Median : 55.0
##  Mean   : 88.64   Mean   : 95.9   Mean   :  2.075   Mean   : 54.6
##  3rd Qu.:122.00   3rd Qu.:132.0   3rd Qu.: 15.000   3rd Qu.: 78.0
##  Max.   :844.00   Max.   :994.0   Max.   : 28.000   Max.   :100.0
##  NA's   :27653    NA's   :2197    NA's   :5         NA's   :339
##       PRES           TEMP          cbwd           Iws
##  Min.   : 991   Min.   :-19.00   cv  :11412   Min.   :  0.45
##  1st Qu.:1008   1st Qu.:  2.00   NE  : 6178   1st Qu.:  1.79
##  Median :1016   Median : 14.00   NW  :16717   Median :  4.92
##  Mean   :1016   Mean   : 12.59   SE  :18272   Mean   : 23.26
##  3rd Qu.:1025   3rd Qu.: 23.00   NA's:    5   3rd Qu.: 21.02
##  Max.   :1046   Max.   : 42.00                Max.   :585.60
##  NA's   :339    NA's   :5                     NA's   :5
##  precipitation          Iprec
##  Min.   :     0.0   Min.   :     0.0
##  1st Qu.:     0.0   1st Qu.:     0.0
##  Median :     0.0   Median :     0.0
##  Mean   :    19.3   Mean   :    19.5
##  3rd Qu.:     0.0   3rd Qu.:     0.0
##  Max.   :999990.0   Max.   :999990.0
##  NA's   :484        NA's   :484
```

```
## columns highted by describe result to detect missing value
## PM_Dongsi,PM_Dongsihuan,PM_Nongzhanguan,PM_US.Post,HUMI,PRES,precipitation,Iprec
```

A.2 Missing Value Visualization

```
mice_plot <- aggr(air, numbers=TRUE, sortVars=TRUE, labels=names(air), cex.axis=0.7, ylab=c("Missing da
```

```
## Warning in plot.aggr(res, ...): not enough vertical space to display
## frequencies (too many combinations)
```




```
##
##  Variables sorted by number of missings:
##       Variable        Count
##     PM_Dongsih 6.099954e-01
##      PM_Nongzh 5.258824e-01
##      PM_Dongsi 5.235813e-01
##     PM_US.Post 4.178077e-02
##  precipitation 9.204321e-03
##          Iprec 9.204321e-03
##           HUMI 6.446828e-03
##           PRES 6.446828e-03
##           DEWP 9.508596e-05
##           TEMP 9.508596e-05
##           cbwd 9.508596e-05
##            Iws 9.508596e-05
##             No 0.000000e+00
##           year 0.000000e+00
##          month 0.000000e+00
##            day 0.000000e+00
##           hour 0.000000e+00
##         season 0.000000e+00
```

The above graph seems little confusing, but it has lot of info.

For instance, the left block shows:

```
- There are 60% missing value in PM_Dongsih
- There are 50% missing value in PM_Nongzh
- There are 50% missing value in PM_Dongsi
- Other columns have less than .5% missing value
```

A.2.1 Before we delve into the nuisances of missign value imputations, lets quickly understand the types of missing value:

- **MCAR** (*Missing Completely at Random*)

It means the missing values are completely random, there is no pattern associated to it. The **Missing** and **Observed** distrubutions will look the same. If we discard these observations, we are not inducing any bias into our analysis. However, one has to be very careful before deleting missing values. If the missing data is less than **0.5%**, then it is safe to delete it, or else one must impute it.

- **MAR** (*Missing at Random*)

It means there is a significant difference between **Missing** and **Observed** distributions, and the pattern can be explained by other observed variables. One way is to run regression individually on each variables (that contains missing value) as response variables, or use the nice **mice** package from R.

- **MNAR** (*Missing not at random*)

This is the most toughest category to impute, generally people delete such columns. For instance, in a survey people with high income left the income column blank. There is no way we can impute such values, because the missing values is dependent on one segment of our income variable, which cannot be imputed. However in such scenarios, it is sometimes advisable to bin the income variable into several buckets, and categorize them into 1 to 5.

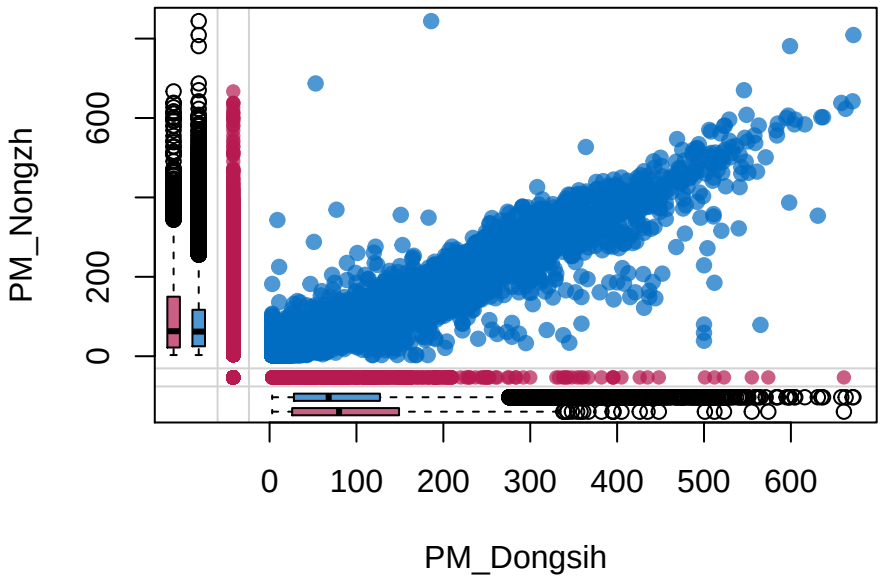
A.3 Understand the missing value pattern

A.3.1 Margin Plot

- It is very helpful in determining if the values are MAR or MCAR
- The margin plot, plots 2 features at a time, and contains box plot & scatter plot

Lets plot PM_Dongsih and PM_Nongzh and interpret the graph

```
marginplot(air[, c("PM_Dongsih", "PM_Nongzh")], col = mdc(1:2), cex.numbers = 1.2, pch = 19)
```



The *red points* displays *missing values* and similarly *blue points* displays *non-missing values*

- Left box plot + Red bar shows the distribution of PM_Nongzh with PM_Dongsih missing (27653 PM_Dongsih) + Blue Bar shows the distribution of PM_Nongzh with observed PM_Donsih + Both the bar shows same mean

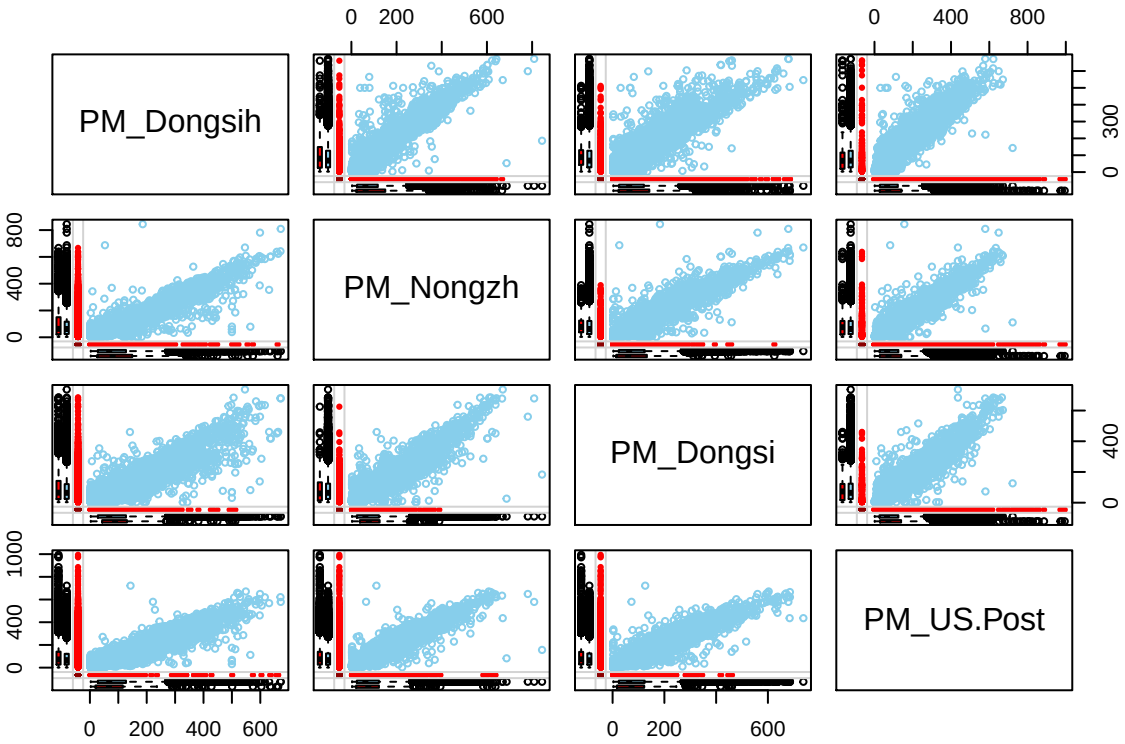
- Right Horizontal box plot
  - Red bar shows the distribution of PM_Dongsih with PM_Nongzh missing (27168 PM_Dongsih)
  - Blue Bar shows the distribution of PM_Dongsih with observed PM_Nongzh
  - The mean is little shifted

- Scatter plot
  - The scatter plot between observed values of PM_Nongzh & PM_Dongsih
  - It shows a linear relationship between the 2 variables. The graph shows that these variables can used as a regressor to predict the missing value pattern amongst them.

A.3.2 Margin Matrix

This plot is simlar to the Margin Plot we discussed above, however one can use multiple variables at a time to give wide perspective.

```
marginmatrix(air[,c("PM_Dongsih", "PM_Nongzh","PM_Dongsi","PM_US.Post")])
```



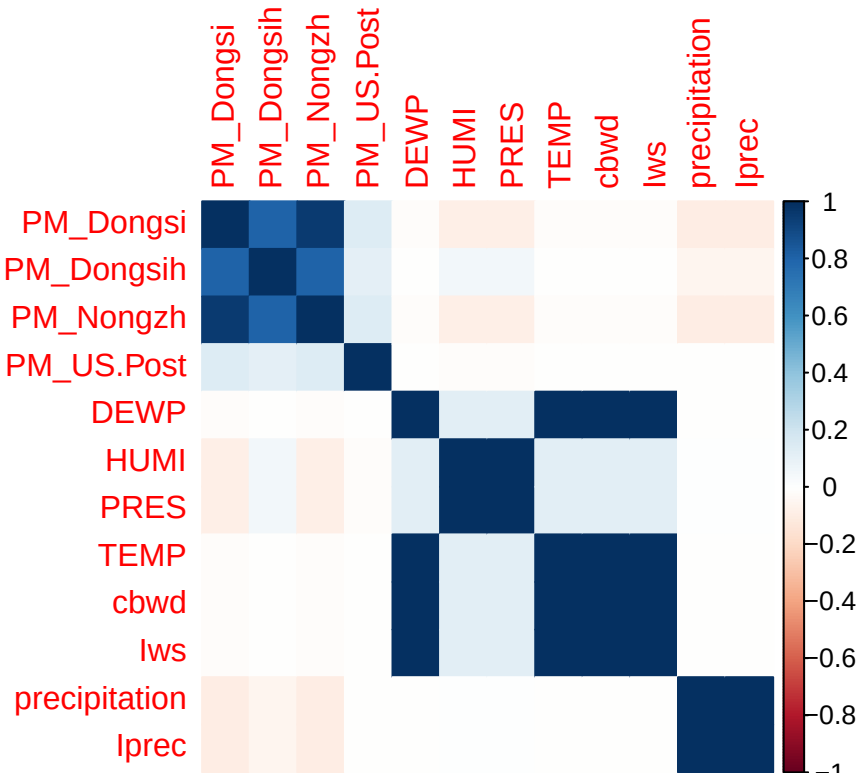
A.3.3 Correlation

We can use correlation to understand the relation between the varibales that have missing values.

- If there is a crorrelation, then MCAR is present.
- If there is no correlation, then MAR is present.
- Unfortunately, there is no test for MNAR (The business domain expert knowledge and the collection process can give us some insights into why it is missing)

```
## 1 if the value in air is missing and 0 if the value is non missing
air.missing <- as.data.frame(abs(is.na(air)))

air.missing.few.col <- air.missing[,sapply(air.missing, sd)>0]
corrplot(cor(air.missing.few.col),method ="color")
```



The corrplot shows extreme correlation between the variables. For example: - PM_Dongsi is correlated to PM_Dongsih & PM_Nongzh - DEWP is correlated to TEMP, cbwd & IWS.

B. Treatment of Missing Value

There are mutiple ways to impute the outliers, we will discuss the most frequnt methods used.

B.1 Delete the observations

If the missing column is less than .5% of the data, then its advisable to delete the observations, as imputing them will induce more bias. Also the effort made to impute using methods discussed below, its not worth the result.

B.2 Delete the variable

This method depends totally on the business knowledge and the expertise one has in the domain. Unless the varibale is so significant (which expalins more than 50% of you model, or very important variable of you analysis), then one can afford to delete the varibale.

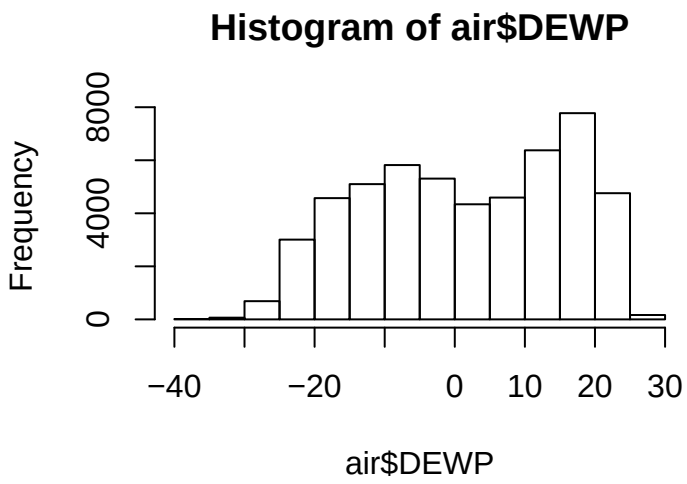
B.3 Imputation by mesaure of central dispersions (mean/median)

This is most generalized, highly effective and unequivocably most used method. Its very easy to use and intuitive.However, I would recommned it touse on varibales, where missing values are less than 1%

The varibales below have missing percentages less than 1%, however one should be careful in choosing mean or median

- If the variable follows a normal distribution (no skewness), then replace by mean.
- If the variable follows a skewed distribution, replace by median.

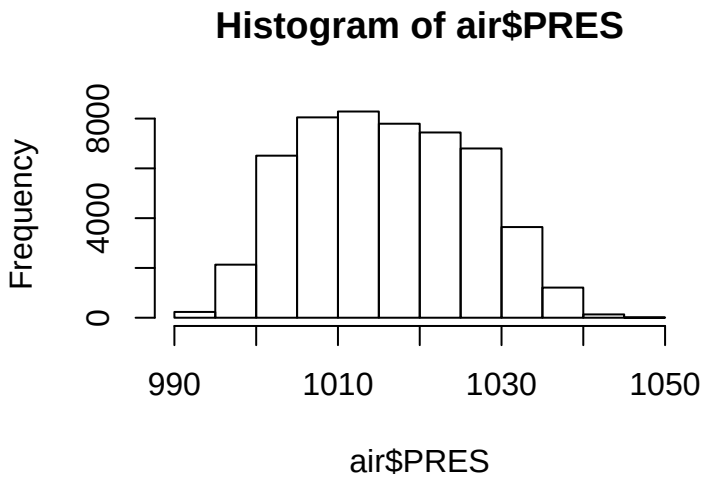
```
## plot the histogram of DEWP
hist(air$DEWP)
```




```
## It is left skewed, hence replace by median
air$DEWP[is.na(air$DEWP)] <- median(air$DEWP, na.rm=TRUE)

## plot the histogram of PRES

hist(air$PRES)
```




```
## It is normal, hence replace by mean
air$PRES[is.na(air$PRES)] <- mean(air$PRES, na.rm=TRUE)


air$HUMI[is.na(air$HUMI)] <- mean(air$HUMI, na.rm=TRUE)

air$TEMP[is.na(air$TEMP)] <- median(air$TEMP, na.rm=TRUE)
```

B.4 MICE Imputation

There has been numerous journals and papers published, and they are cited as reference, in case one wants to dive into the depth. I will be touching the preliminary concepts required:

B.4.1 MICE algorithm

MICE Multiple Imputation by Chained Equations

Step 1 : Impute all the missing values, think of it has "placeholders"

Step 2 : Pick one var1, remove the placeholders, var1 has both *observed* and *missing*

Step 3 : Run regression on observed values of **var1** as response variable, and other variables as predictors(sometimes all or subset of the variables are used for the prediction).

Step 4 : Use the above model to predict the missing values. - When the above **var1** is used as independent variable, both the observed and imputed values will be used to train the model.

Step 5 : Step 2 to Step 4 is repeated for every variable which has missing values.

Step 6: The above steps run for mutiple iterations, as in Step 3 not all the variables have been picked. It tries to run many iterations to store as much imouted values it can, just like radnom forest works. Do boot strapping and estimate as much as trees as we want, so that we all the pssible combinations of variables.

B.4.2 MICE Imputation in R

```
set.seed(1234)
air.imputed.mice <- mice(air,m=1, maxit=3, meth="pmm")
```

```
##
##  iter imp variable
##   1   1  PM_Dongsi  PM_Dongsih  PM_Nongzh  PM_US.Post  cbwd  Iws  precipitation
##   2   1  PM_Dongsi  PM_Dongsih  PM_Nongzh  PM_US.Post  cbwd  Iws  precipitation
##   3   1  PM_Dongsi  PM_Dongsih  PM_Nongzh  PM_US.Post  cbwd  Iws  precipitation
```

```
#summary(air.imputed.mice)
```

Lets see what does all the parameter means:

- **m** : number of imputed datasets
- **maxit**: number of iterations (cycles from Step 2- Step 4) one wants to run (generally keep it low, when the dataset is big)
- **meth**: meth stands for method, i.e. there are multiple methods available
  - *pmm*: predictive mean matching
  - *rf*: rando forest imputation
  - *cart*: classification or regression trees

We can run summary of the data set just obtained, however its little difficult to read it, hence omitted that step in output.

B.4.3 MICE Imputation complete function

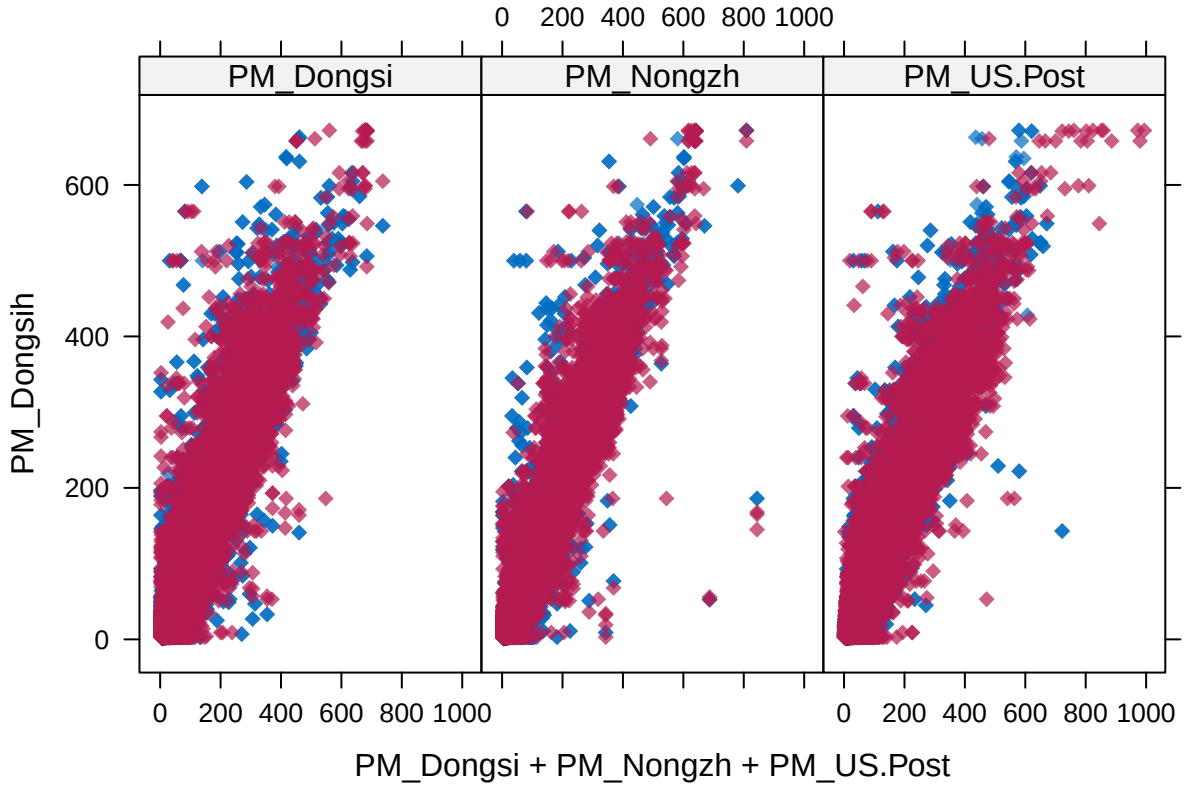
We can use **complete** function to call which imputed data set we want to use

- if we want **1st data set**, use complete with **parameter 1**
- if we want the **2nd data set**, use complete with **paramter 2**

B.4.4 MICE Imputation distribution check

Its nice to check how the observed and imouted values looks like:

```
xyplot(air.imputed.mice,PM_Dongsih ~ PM_Dongsi+PM_Nongzh+PM_US.Post,pch=18,cex=1)
```



The red dots represents the imputed values, while the blue dots represnts the observed values. From the graph, it is clear that imputed values have same shape as of observed values with respect to other variables.

Thank you for reading the article.

The first 3 papers mentioned in the references are great, and I highly recommned it to read if someone wants to master the art of imputation.

References

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/
2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4121561/
3. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4701517/
4. http://r-statistics.co/Missing-Value-Treatment-With-R.html
5. https://datascienceplus.com/imputing-missing-data-with-r-mice-package/