Impute evaluation vignette

## Package install and data import

library(magrittr)  
library(knitr)  
library(ggplot2)  
library(reshape2)  
source('Imputation evaluations.R')  
data\_test <- read.csv('OB\_data/Real\_data\_DM.csv', row.names = 1)  
group <- rownames(data\_test) %>% gsub('()-.\*', '\\1', .) %>% as.factor()

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | var\_1 | var\_2 | var\_3 | var\_4 | var\_5 | var\_6 | var\_7 | var\_8 |
| DM-1 | 70.791 | 514.177 | 206.219 | 411.484 | 4.721 | 21.626 | 7.300 | 0.648 |
| DM-10 | 50.623 | 128.610 | 41.453 | 227.880 | 0.540 | 47.970 | 4.781 | 2.123 |
| DM-11 | 42.720 | 125.704 | 46.343 | 160.441 | 2.990 | 9.155 | 9.994 | 0.524 |
| DM-12 | 15.682 | 67.553 | 21.916 | 71.591 | 0.728 | 23.039 | 5.737 | 2.506 |
| DM-13 | 65.920 | 72.615 | 18.242 | 123.881 | 1.381 | 18.932 | 3.326 | 0.708 |
| DM-14 | 161.750 | 379.706 | 26.088 | 274.699 | 0.490 | 43.720 | 15.053 | 4.504 |

## [1] DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM  
## [24] DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM  
## [47] DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM DM  
## [70] DM N N N N N N N N N N N N N N N N N N N N N N   
## [93] N N N N N N N N N N N N N N N N N N N N N N N   
## [116] N N N N N N N N N N N N N N N N N N N N N N N   
## [139] N N N N N N N N N N N N N N N N N N N N N N N   
## [162] N N N N N N N N N N N N N N N N N N N N N N N   
## [185] N N N N N N N N N N N N N N   
## Levels: DM N

# MCAR

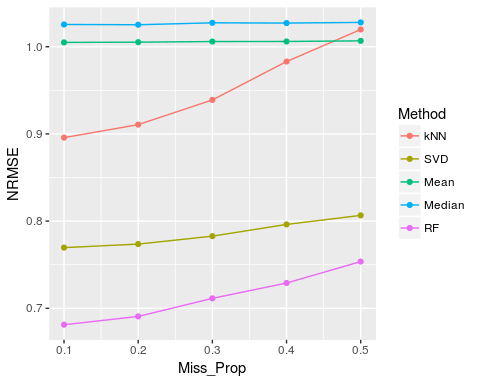
## MCAR generation and imputation

MCAR\_list <- MCAR\_gen\_imp(data\_c = data\_test, prop = seq(.1, .5, .1), impute\_list = c('kNN\_wrapper', 'SVD\_wrapper', 'Mean\_wrapper', 'Median\_wrapper', 'RF\_wrapper'), cores = 10)

## MCAR NRMSE evaluation and plot

MCAR\_NRMSE\_list <- NRMSE\_cal\_plot(MCAR\_list, plot = T, x = 'Miss\_Prop')

## [1] 1  
## [1] 2  
## [1] 3  
## [1] 4  
## [1] 5



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | Mean | Median | RF | Miss\_Prop | Miss\_Num |
| 0.8957382 | 0.7696032 | 1.004958 | 1.025544 | 0.6811728 | 0.1 | 130 |
| 0.9106655 | 0.7736548 | 1.005224 | 1.025213 | 0.6906557 | 0.2 | 130 |
| 0.9389917 | 0.7827661 | 1.005945 | 1.027410 | 0.7113980 | 0.3 | 130 |
| 0.9830233 | 0.7961935 | 1.006039 | 1.027144 | 0.7289833 | 0.4 | 130 |
| 1.0197809 | 0.8065801 | 1.006747 | 1.027885 | 0.7535703 | 0.5 | 130 |

## The above table shows the NRMSE of different imputaion methods

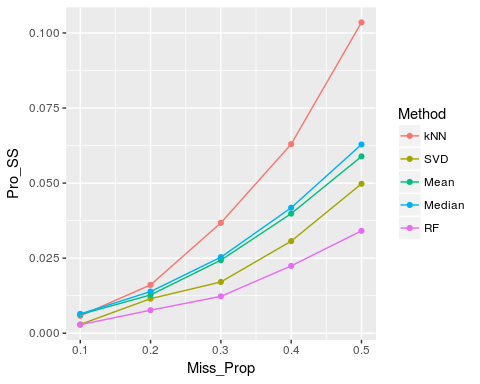
|  |  |  |
| --- | --- | --- |
| Miss\_Prop | Method | NRMSE |
| 0.1 | kNN | 0.8957382 |
| 0.2 | kNN | 0.9106655 |
| 0.3 | kNN | 0.9389917 |
| 0.4 | kNN | 0.9830233 |
| 0.5 | kNN | 1.0197809 |
| 0.1 | SVD | 0.7696032 |

## The above melted table is good for ggplot2

## MCAR PCA Procrustes analysis and plot

MCAR\_PCA\_ProSS\_list <- Procrustes\_cal\_plot(MCAR\_list, DR = 'PCA', nPCs = 2, x = 'Miss\_Prop', plot = T)

## [1] 1  
## [1] 2  
## [1] 3  
## [1] 4  
## [1] 5



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | Mean | Median | RF | Miss\_Prop | Miss\_Num |
| 0.0058081 | 0.0029227 | 0.0062499 | 0.0064181 | 0.0028319 | 0.1 | 130 |
| 0.0160971 | 0.0115052 | 0.0127537 | 0.0138729 | 0.0076673 | 0.2 | 130 |
| 0.0367555 | 0.0170723 | 0.0243381 | 0.0253273 | 0.0122743 | 0.3 | 130 |
| 0.0629755 | 0.0306623 | 0.0398333 | 0.0417957 | 0.0224246 | 0.4 | 130 |
| 0.1035155 | 0.0497668 | 0.0588806 | 0.0628540 | 0.0340962 | 0.5 | 130 |

## The above table shows the Procrustes Sum of Squared Error of different imputaion methods

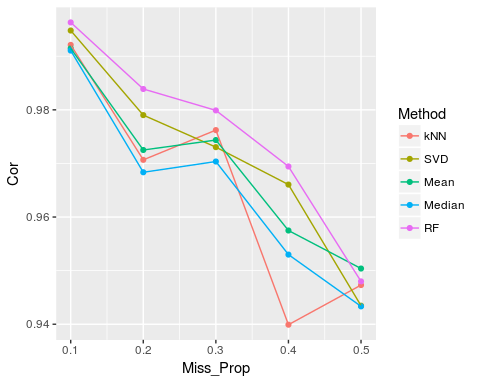
|  |  |  |
| --- | --- | --- |
| Miss\_Prop | Method | Pro\_SS |
| 0.1 | kNN | 0.0058081 |
| 0.2 | kNN | 0.0160971 |
| 0.3 | kNN | 0.0367555 |
| 0.4 | kNN | 0.0629755 |
| 0.5 | kNN | 0.1035155 |
| 0.1 | SVD | 0.0029227 |

## The above melted table is good for ggplot2

## MCAR T-test results correlation

MCAR\_Ttest\_cor\_list <- Ttest\_cor\_cal\_plot(MCAR\_list, group = group, plot = T, x = 'Miss\_Prop', cor = 'P')

## [1] 1  
## [1] 2  
## [1] 3  
## [1] 4  
## [1] 5



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | Mean | Median | RF | Miss\_Prop | Miss\_Num |
| 0.9921166 | 0.9947967 | 0.9914626 | 0.9910585 | 0.9963095 | 0.1 | 130 |
| 0.9706710 | 0.9790125 | 0.9725072 | 0.9683502 | 0.9838864 | 0.2 | 130 |
| 0.9762177 | 0.9730374 | 0.9743628 | 0.9703540 | 0.9798917 | 0.3 | 130 |
| 0.9399089 | 0.9660529 | 0.9574841 | 0.9529879 | 0.9694339 | 0.4 | 130 |
| 0.9473047 | 0.9434554 | 0.9503873 | 0.9433293 | 0.9479859 | 0.5 | 130 |

## The above table shows the Pearson Correlation of log T-test P-values between imputed data and complete data

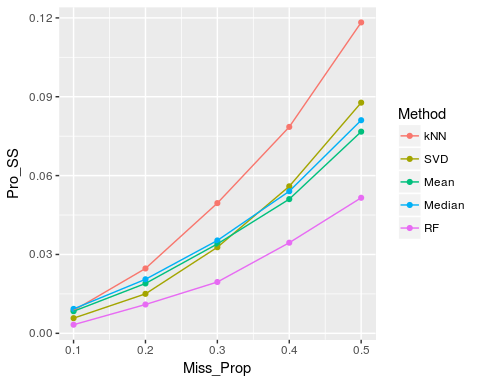
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | Mean | Median | RF | Miss\_Prop | Miss\_Num |
| 0.9906485 | 0.9914788 | 0.9890754 | 0.9883762 | 0.9939915 | 0.1 | 130 |
| 0.9803247 | 0.9781016 | 0.9776427 | 0.9761952 | 0.9877699 | 0.2 | 130 |
| 0.9700610 | 0.9571263 | 0.9714758 | 0.9686299 | 0.9732510 | 0.3 | 130 |
| 0.9372598 | 0.9338513 | 0.9505933 | 0.9454861 | 0.9560120 | 0.4 | 130 |
| 0.9098718 | 0.9009081 | 0.9193762 | 0.9063868 | 0.9426621 | 0.5 | 130 |

## The above table shows the Spearman Correlation of T-test P-values between imputed data and complete data

## MCAR PLS Procrustes analysis and plot

MCAR\_PLS\_ProSS\_list <- Procrustes\_cal\_plot(MCAR\_list, DR = 'PLS', nPCs = 2, outcome = group, x = 'Miss\_Prop', plot = T)

## [1] 1  
## [1] 2  
## [1] 3  
## [1] 4  
## [1] 5



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | Mean | Median | RF | Miss\_Prop | Miss\_Num |
| 0.0086773 | 0.0057721 | 0.0084034 | 0.0092919 | 0.0032826 | 0.1 | 130 |
| 0.0246732 | 0.0150190 | 0.0189715 | 0.0205573 | 0.0109688 | 0.2 | 130 |
| 0.0495326 | 0.0327509 | 0.0340160 | 0.0353257 | 0.0195211 | 0.3 | 130 |
| 0.0785230 | 0.0559197 | 0.0510504 | 0.0540521 | 0.0344740 | 0.4 | 130 |
| 0.1182722 | 0.0877537 | 0.0767151 | 0.0810655 | 0.0515663 | 0.5 | 130 |

## The above table shows the Procrustes Sum of Squared Error of different imputaion methods

|  |  |  |
| --- | --- | --- |
| Miss\_Prop | Method | Pro\_SS |
| 0.1 | kNN | 0.0086773 |
| 0.2 | kNN | 0.0246732 |
| 0.3 | kNN | 0.0495326 |
| 0.4 | kNN | 0.0785230 |
| 0.5 | kNN | 0.1182722 |
| 0.1 | SVD | 0.0057721 |

# ## The above melted table is good for ggplot2

# MNAR

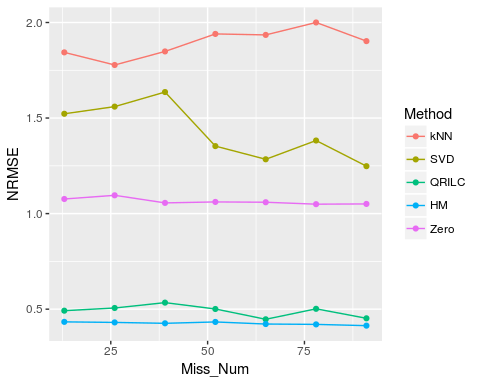
## MNAR generation and imputation

MNAR\_list <- MNAR\_gen\_imp(data\_c = data\_test, mis\_var\_prop = seq(.1, .7, .1), var\_mis\_prop = seq(.3, .6, .1), impute\_list = c('kNN\_wrapper', 'SVD\_wrapper', 'QRILC\_wrapper', 'HM\_wrapper', 'Zero\_wrapper'), cores = 1)

## MNAR NRMSE evaluation and plot

MNAR\_NRMSE\_list <- NRMSE\_cal\_plot(MNAR\_list, plot = T, x = 'Miss\_Num', sc=F)

## [1] 1  
## [1] 2  
## [1] 3  
## [1] 4  
## [1] 5  
## [1] 6  
## [1] 7



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | QRILC | HM | Zero | Miss\_Prop | Miss\_Num |
| 1.843693 | 1.522403 | 0.4919585 | 0.4334591 | 1.076480 | 0.0493007 | 13 |
| 1.777751 | 1.559842 | 0.5060922 | 0.4305481 | 1.095724 | 0.0870629 | 26 |
| 1.848609 | 1.636142 | 0.5341109 | 0.4258740 | 1.056053 | 0.1394328 | 39 |
| 1.940268 | 1.353173 | 0.5008561 | 0.4330214 | 1.061102 | 0.1832945 | 52 |
| 1.935096 | 1.284224 | 0.4469191 | 0.4221009 | 1.059489 | 0.2303030 | 65 |
| 2.000223 | 1.382448 | 0.5017534 | 0.4202394 | 1.049319 | 0.2719503 | 78 |

## The above table shows the NRMSE of different imputaion methods

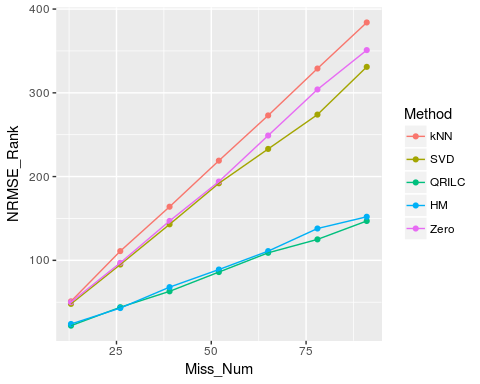
|  |  |  |
| --- | --- | --- |
| Miss\_Num | Method | NRMSE |
| 13 | kNN | 1.843693 |
| 26 | kNN | 1.777751 |
| 39 | kNN | 1.848609 |
| 52 | kNN | 1.940268 |
| 65 | kNN | 1.935096 |
| 78 | kNN | 2.000223 |

## The above melted table is good for ggplot2

## MNAR NRMSE rank evaluation and plot

MNAR\_NRMSE\_rank\_list <- NRMSE\_rank\_cal\_plot(MNAR\_list, plot = T, x = 'Miss\_Num')

## [1] 1  
## [1] 2  
## [1] 3  
## [1] 4  
## [1] 5  
## [1] 6  
## [1] 7



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | QRILC | HM | Zero | Miss\_Prop | Miss\_Num |
| 51 | 48 | 22 | 24 | 50 | 0.0493007 | 13 |
| 111 | 95 | 44 | 43 | 97 | 0.0870629 | 26 |
| 164 | 143 | 63 | 68 | 147 | 0.1394328 | 39 |
| 219 | 192 | 86 | 89 | 194 | 0.1832945 | 52 |
| 273 | 233 | 109 | 111 | 249 | 0.2303030 | 65 |
| 329 | 274 | 125 | 138 | 304 | 0.2719503 | 78 |

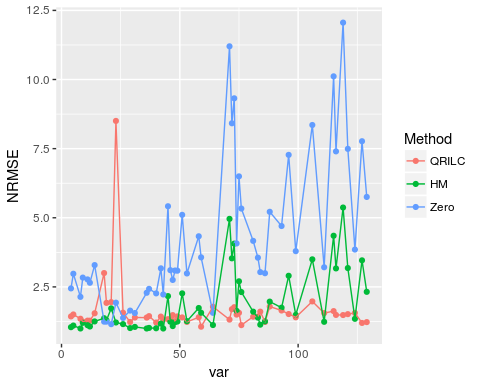
## The above table shows the NRMSE ranks of different imputaion methods

|  |  |  |
| --- | --- | --- |
| Miss\_Num | Method | NRMSE\_Rank |
| 13 | kNN | 51 |
| 26 | kNN | 111 |
| 39 | kNN | 164 |
| 52 | kNN | 219 |
| 65 | kNN | 273 |
| 78 | kNN | 329 |

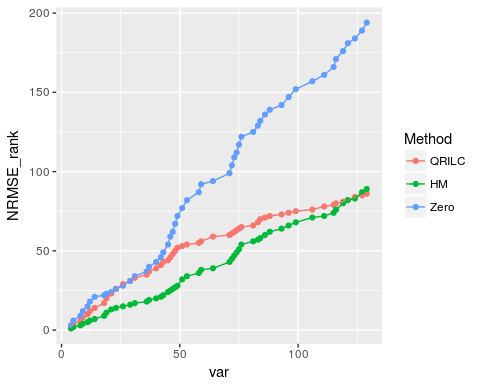
## The above melted table is good for ggplot2

## NRMSE and SOR on MNAR

## We randomly choose a cut point of 4



## NRMSE of three MNAR imputation methods for each missing variable. Zero showed overall unstable and poor performances. QRILC showed overall stable and good performances except for one missing variable.

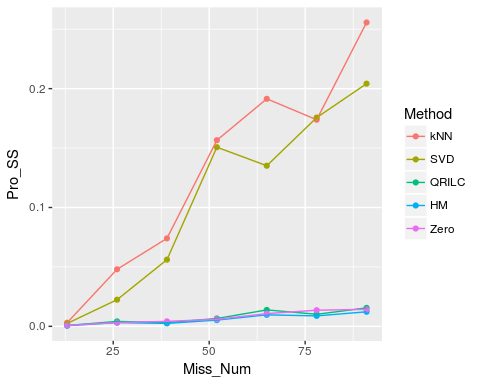


## Cumulative ranks of three MNAR imputation methods for each missing variable. It’s a more robust measurement compared with NRMSE.

## MNAR PCA Procrustes analysis and plot

MNAR\_PCA\_ProSS\_list <- Procrustes\_cal\_plot(MNAR\_list, DR = 'PCA', nPCs = 2, x = 'Miss\_Num', plot = T)

## [1] 1  
## [1] 2  
## [1] 3  
## [1] 4  
## [1] 5  
## [1] 6  
## [1] 7



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | QRILC | HM | Zero | Miss\_Prop | Miss\_Num |
| 0.0030569 | 0.0024599 | 0.0007304 | 0.0005824 | 0.0008678 | 0.0493007 | 13 |
| 0.0480066 | 0.0224202 | 0.0041478 | 0.0031385 | 0.0031507 | 0.0870629 | 26 |
| 0.0740282 | 0.0561170 | 0.0031186 | 0.0024640 | 0.0042093 | 0.1394328 | 39 |
| 0.1565980 | 0.1507121 | 0.0066117 | 0.0052507 | 0.0059363 | 0.1832945 | 52 |
| 0.1914342 | 0.1351302 | 0.0138310 | 0.0096825 | 0.0107177 | 0.2303030 | 65 |
| 0.1738227 | 0.1756455 | 0.0101566 | 0.0088044 | 0.0135821 | 0.2719503 | 78 |

## The above table shows the Procrustes Sum of Squared Error of different imputaion methods

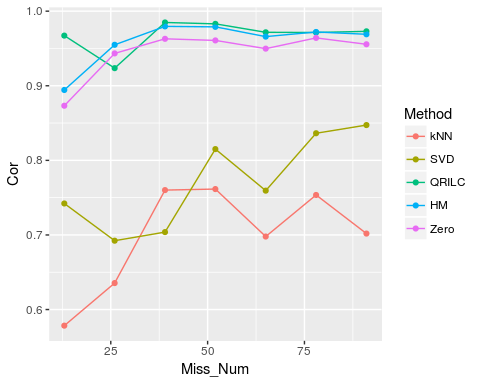
|  |  |  |
| --- | --- | --- |
| Miss\_Num | Method | Pro\_SS |
| 13 | kNN | 0.0030569 |
| 26 | kNN | 0.0480066 |
| 39 | kNN | 0.0740282 |
| 52 | kNN | 0.1565980 |
| 65 | kNN | 0.1914342 |
| 78 | kNN | 0.1738227 |

## The above melted table is good for ggplot2

## MNAR T-test results correlation

MNAR\_Ttest\_cor\_list <- Ttest\_cor\_cal\_plot(MNAR\_list, group = group, plot = T, x = 'Miss\_Num', cor = 'P')

## [1] 1  
## [1] 2  
## [1] 3  
## [1] 4  
## [1] 5  
## [1] 6  
## [1] 7



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | QRILC | HM | Zero | Miss\_Prop | Miss\_Num |
| 0.5784013 | 0.7421768 | 0.9672143 | 0.8943420 | 0.8732141 | 0.0493007 | 13 |
| 0.6354761 | 0.6922887 | 0.9235852 | 0.9549122 | 0.9433824 | 0.0870629 | 26 |
| 0.7601306 | 0.7038311 | 0.9848421 | 0.9795935 | 0.9629277 | 0.1394328 | 39 |
| 0.7615114 | 0.8150958 | 0.9829798 | 0.9789519 | 0.9607697 | 0.1832945 | 52 |
| 0.6978657 | 0.7593904 | 0.9716515 | 0.9658767 | 0.9497483 | 0.2303030 | 65 |
| 0.7535756 | 0.8363294 | 0.9713824 | 0.9720855 | 0.9641362 | 0.2719503 | 78 |

## The above table shows the Pearson Correlation of log T-test P-values between imputed data and complete data

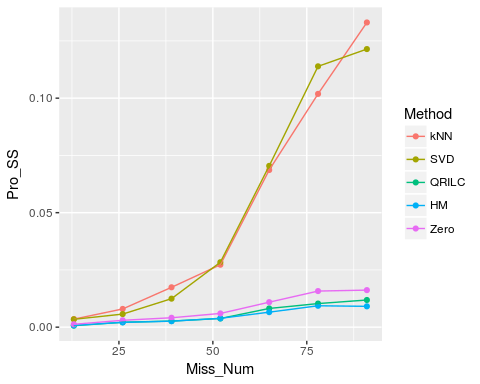
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | QRILC | HM | Zero | Miss\_Prop | Miss\_Num |
| 0.3901099 | 0.2362637 | 0.9285714 | 0.9725275 | 0.9505495 | 0.0493007 | 13 |
| 0.6348718 | 0.5829060 | 0.9589744 | 0.9254701 | 0.9295726 | 0.0870629 | 26 |
| 0.7064777 | 0.6530364 | 0.9777328 | 0.9672065 | 0.9506073 | 0.1394328 | 39 |
| 0.5924187 | 0.6535473 | 0.9692649 | 0.9651669 | 0.9531290 | 0.1832945 | 52 |
| 0.5629808 | 0.6408217 | 0.9640297 | 0.9696241 | 0.9615385 | 0.2303030 | 65 |
| 0.6645380 | 0.7563702 | 0.9702070 | 0.9765551 | 0.9713957 | 0.2719503 | 78 |

## The above table shows the Spearman Correlation of T-test P-values between imputed data and complete data

## MNAR PLS Procrustes analysis and plot

MNAR\_PLS\_ProSS\_list <- Procrustes\_cal\_plot(MNAR\_list, DR = 'PLS', nPCs = 2, outcome = group, x = 'Miss\_Num', plot = T)

## [1] 1  
## [1] 2  
## [1] 3  
## [1] 4  
## [1] 5  
## [1] 6  
## [1] 7



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| kNN | SVD | QRILC | HM | Zero | Miss\_Prop | Miss\_Num |
| 0.0035220 | 0.0034429 | 0.0006686 | 0.0008045 | 0.0012759 | 0.0493007 | 13 |
| 0.0079679 | 0.0057215 | 0.0021548 | 0.0021177 | 0.0030167 | 0.0870629 | 26 |
| 0.0174182 | 0.0124557 | 0.0026718 | 0.0026206 | 0.0041173 | 0.1394328 | 39 |
| 0.0272772 | 0.0283593 | 0.0037994 | 0.0038734 | 0.0060006 | 0.1832945 | 52 |
| 0.0686296 | 0.0704292 | 0.0081831 | 0.0065649 | 0.0109262 | 0.2303030 | 65 |
| 0.1018619 | 0.1138955 | 0.0103098 | 0.0093510 | 0.0157715 | 0.2719503 | 78 |

## The above table shows the Procrustes Sum of Squared Error of different imputaion methods

|  |  |  |
| --- | --- | --- |
| Miss\_Num | Method | Pro\_SS |
| 13 | kNN | 0.0035220 |
| 26 | kNN | 0.0079679 |
| 39 | kNN | 0.0174182 |
| 52 | kNN | 0.0272772 |
| 65 | kNN | 0.0686296 |
| 78 | kNN | 0.1018619 |

## The above melted table is good for ggplot2