Tidy Data

Sandeep Joshi

6/4/2019

### Missing Values

library(tidyverse)  
library(readr)  
library(knitr)  
library(Hmisc)  
library(missForest)

## Missing Values

Set missing values ( 10% )

iris.mis <- prodNA(iris, noNA = 0.1)  
kable(summary(iris.mis))

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
|  | Min. :4.300 | Min. :2.000 | Min. :1.000 | Min. :0.100 | setosa :45 |
|  | 1st Qu.:5.100 | 1st Qu.:2.800 | 1st Qu.:1.600 | 1st Qu.:0.300 | versicolor:44 |
|  | Median :5.800 | Median :3.000 | Median :4.400 | Median :1.300 | virginica :45 |
|  | Mean :5.843 | Mean :3.049 | Mean :3.786 | Mean :1.176 | NA’s :16 |
|  | 3rd Qu.:6.400 | 3rd Qu.:3.350 | 3rd Qu.:5.100 | 3rd Qu.:1.800 | NA |
|  | Max. :7.900 | Max. :4.400 | Max. :6.900 | Max. :2.500 | NA |
|  | NA’s :12 | NA’s :15 | NA’s :16 | NA’s :16 | NA |

print(paste0("Missing Values:", sum(is.na(iris.mis)==TRUE)))

## [1] "Missing Values:75"

Impute with mean value

iris.mis$imputed\_age <- with(iris.mis, impute(Sepal.Length, mean))

Impute using argImpute

impute\_arg <- aregImpute(~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width +  
Species, data = iris.mis, n.impute = 5)

## Iteration 1   
Iteration 2   
Iteration 3   
Iteration 4   
Iteration 5   
Iteration 6   
Iteration 7   
Iteration 8

impute\_arg

##   
## Multiple Imputation using Bootstrap and PMM  
##   
## aregImpute(formula = ~Sepal.Length + Sepal.Width + Petal.Length +   
## Petal.Width + Species, data = iris.mis, n.impute = 5)  
##   
## n: 150 p: 5 Imputations: 5 nk: 3   
##   
## Number of NAs:  
## Sepal.Length Sepal.Width Petal.Length Petal.Width Species   
## 12 15 16 16 16   
##   
## type d.f.  
## Sepal.Length s 2  
## Sepal.Width s 2  
## Petal.Length s 2  
## Petal.Width s 2  
## Species c 2  
##   
## Transformation of Target Variables Forced to be Linear  
##   
## R-squares for Predicting Non-Missing Values for Each Variable  
## Using Last Imputations of Predictors  
## Sepal.Length Sepal.Width Petal.Length Petal.Width Species   
## 0.867 0.725 0.982 0.952 0.985

impute\_arg$imputed$Sepal.Length # check

## [,1] [,2] [,3] [,4] [,5]  
## 18 5.0 5.0 5.0 5.2 5.1  
## 20 5.4 4.8 5.4 5.0 5.1  
## 23 4.9 5.0 5.2 5.0 5.4  
## 32 5.4 4.9 5.2 5.1 5.2  
## 37 5.2 4.6 5.1 5.1 5.0  
## 52 6.3 6.7 6.4 6.4 6.4  
## 66 6.0 6.5 6.4 5.9 6.6  
## 82 5.6 5.6 5.1 4.8 5.1  
## 109 6.0 6.8 7.1 6.5 6.3  
## 111 6.9 6.8 5.9 6.5 6.8  
## 129 6.0 6.3 6.4 6.7 6.5  
## 149 6.8 6.8 6.3 6.9 6.4

religion = read\_csv("religion\_income.csv")

## Gather data from wide to long table

kable(religion)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| religion | <$10k | $10–20k | $20–30k | $30–40k | $40–50k | $50–75k | $75–100k | $100–150k | >150k | Don’t know/refused |
| Agnostic | 27 | 34 | 60 | 81 | 76 | 137 | 122 | 109 | 84 | 96 |
| Atheist | 12 | 27 | 37 | 52 | 35 | 70 | 73 | 59 | 74 | 76 |
| Buddhist | 27 | 21 | 30 | 34 | 33 | 58 | 62 | 39 | 53 | 54 |
| Catholic | 418 | 617 | 732 | 670 | 638 | 1116 | 949 | 792 | 633 | 1489 |
| Don’t know/refused | 15 | 14 | 15 | 11 | 10 | 35 | 21 | 17 | 18 | 116 |
| Evangelical Prot | 575 | 869 | 1064 | 982 | 881 | 1486 | 949 | 723 | 414 | 1529 |
| Hindu | 1 | 9 | 7 | 9 | 11 | 34 | 47 | 48 | 54 | 37 |
| Historically Black Prot | 228 | 244 | 236 | 238 | 197 | 223 | 131 | 81 | 78 | 339 |
| Jehovah’s Witness | 20 | 27 | 24 | 24 | 21 | 30 | 15 | 11 | 6 | 37 |
| Jewish | 19 | 19 | 25 | 25 | 30 | 95 | 69 | 87 | 151 | 162 |
| Mainline Prot | 289 | 495 | 619 | 655 | 651 | 1107 | 939 | 753 | 634 | 1328 |
| Mormon | 29 | 40 | 48 | 51 | 56 | 112 | 85 | 49 | 42 | 69 |
| Muslim | 6 | 7 | 9 | 10 | 9 | 23 | 16 | 8 | 6 | 22 |
| Orthodox | 13 | 17 | 23 | 32 | 32 | 47 | 38 | 42 | 46 | 73 |
| Other Christian | 9 | 7 | 11 | 13 | 13 | 14 | 18 | 14 | 12 | 18 |
| Other Faiths | 20 | 33 | 40 | 46 | 49 | 63 | 46 | 40 | 41 | 71 |
| Other World Religions | 5 | 2 | 3 | 4 | 2 | 7 | 3 | 4 | 4 | 8 |
| Unaffiliated | 217 | 299 | 374 | 365 | 341 | 528 | 407 | 321 | 258 | 597 |

long <- religion %>%   
 gather(income, freq, 2:11)  
kable(long[1:10,])

|  |  |  |
| --- | --- | --- |
| religion | income | freq |
| Agnostic | <$10k | 27 |
| Atheist | <$10k | 12 |
| Buddhist | <$10k | 27 |
| Catholic | <$10k | 418 |
| Don’t know/refused | <$10k | 15 |
| Evangelical Prot | <$10k | 575 |
| Hindu | <$10k | 1 |
| Historically Black Prot | <$10k | 228 |
| Jehovah’s Witness | <$10k | 20 |
| Jewish | <$10k | 19 |