Data Cleaning-01 Missing Values

Created by H. M. Samadhi Chathuranga Rathnayake

```r
setwd("D:\\Workshops\\R Programming for Data Science Workshop\\Part 02 - Data
Manipulation & Cleaning\\Datasets")

data=read.csv("iris - Missing.CSV")
head(data)

str(data)

data$Species[data$Species==""]=NA
data$Species=factor(data$Species)
str(data)

#Missing values are represented as NA s
#Checking missing values
summary(data)

sum(is.na(data$Petal.Width))/length(data$Petal.Width)

#To view the percentages of missing values in each column
fun=function(x){return(sum(is.na(x))/length(x))}
apply(data,2,fun)

#install.packages("mice")
library(mice)

md.pattern(data)

md.pairs(data)

#Dealing with missing values
data$Petal.Width=NULL #Since this is containing more than 70% of missing
values

summary(data)

#Filling missing values with suitable values
data$Sepal.Length[is.na(data$Sepal.Length)]=mean(data$Sepal.Length,na.rm =
TRUE)
data$Sepal.Length

summary(data)

data$Sepal.Width[is.na(data$Sepal.Width)]=mean(data$Sepal.Width,na.rm = TRUE)
data$Petal.Length[is.na(data$Petal.Length)]=mean(data$Petal.Length,na.rm =
TRUE)
summary(data)
```

```r
table(data$Species)

#names(which.max(table(data$Species)))
#data$Species[is.na(data$Species)]=names(which.max(table(data$Species)))


is.na(data)

nrow(data)

new_data=data[!is.na(data$Species),]
new_data

complete.cases(data)

new_data2=data[complete.cases(data),]
new_data2

#Handling missing values with tidyr
#install.packages("tidyr")
library(tidyr)
setwd("D:\\Workshops\\R Programming for Data Science Workshop\\Part 02 - Data Manipulation & Cleaning\\Datasets")

data=read.csv("iris - Missing.CSV")
head(data)

str(data)

data$Species[data$Species==""]=NA
data$Species=factor(data$Species)
str(data)

#Removing missing value rows
new_data=drop_na(data)
new_data

#Filling missing values
new_data1=fill(data,"Sepal.Length",.direction = "down")
head(new_data1)

new_data2=fill(data,"Sepal.Length","Sepal.Width",.direction = "down")
head(new_data2)

new_data3=fill(data,"Sepal.Length","Sepal.Width",.direction = "up")
head(new_data3)

new_data4=fill(data,"Sepal.Length","Sepal.Width",.direction = "downup")
head(new_data4)

new_data5=fill(data,"Sepal.Length","Sepal.Width",.direction = "updown")
head(new_data5)
```

```
#Replacing missing values
new_data1=replace_na(data,list(Sepal.Length=0))
head(new_data1)

new_data2=replace_na(data,list(Sepal.Length=0,Sepal.Width=0))
head(new_data2)

new_data3=replace_na(data,list(Sepal.Length=mean(data$Sepal.Length,na.rm =
TRUE)))
head(new_data3)
```