Data Manipulation-01 Handling Data in Base R

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.CSV")

#If we want to import an excel file we have to use xlsx library
#install.packages("readxl")
library(readxl)
data_xl=read_xlsx("iris.XLSX")
head(data_xl)

#Top and bottom values
head(data)

tail(data)

#Data types
str(data)

#Dimensions
dim(data)

nrow(data)

ncol(data)

#Columns and indexes
colnames(data)

row.names(data)

#Changing columns or indexes
colnames(data)=c("ID","SL","SW","PL","PW","Spec")
head(data)

#Changing the order of columns
data2=data[,c(1,2,4,3,5,6)]
head(data2)

#Accessing columns and rows
data$SL

data[,1]

data["SL"]

data[c("SL","PL")]
```

```r
data[1,2]

#Removing & adding new columns with column operations
data$ID=NULL
head(data)

data$MeanL=(data$SL+data$PL)/2
data$MeanW=(data$SW+data$PW)/2
head(data)

data$LLevel=ifelse(data$MeanL>=mean(data$MeanL),"High","Low")
data$WLevel=ifelse(data$MeanW>=mean(data$MeanW),"High","Low")
head(data)

#Using apply function
data_num=data[c("SL","SW","PL","PW")]
head(data_num)

apply(data_num, 2, mean)

apply(data_num, 2, function(x) return(x*2))

#Selecting rows that have certain values
data$Spec=="setosa"

dSetosa=data[data$Spec=="setosa",]
head(dSetosa)

data$SL<=5

dSL=data[data$SL<=5,]
dSL

data[apply(data["Spec"], 2, function(x) return(x=="setosa")),]

#Selecting rows that are in a vector
head(data)

vec=c("setosa","versicolor")
data_new=data[data$Spec%in%vec,]
data_new

#Data summaries
summary(data)

rowSums(data_num)

colSums(data_num)

rowMeans(data_num)
```

```r
colMeans(data_num)

apply(data_num, 2, mean)

apply(data_num, 1, mean)

#Any vector function supports data frame columns
sum(data$SL)

mean(data$PL)

median(data$SL)

var(data$SW)

sd(data$PW)

quantile(data$SL)

max(data$SL)

min(data$SL)

#Sorting with a column
order(data$SL)

data_OSL=data[order(data$SL),]
data_OSL

nrow(data_OSL)

row.names(data_OSL)=1:150
data_OSL

order(-data$SL)

data_OSL2=data[order(-data$SL),]
data_OSL2

order(data$SL,data$PL)

data_OSL3=data[order(data$SL,data$PL),]
data_OSL3

#Grouping with a categorical variable
groups=list(data[data$Spec=="setosa",],data[data$Spec=="versicolor",],data[da
ta$Spec=="virginica",])
names(groups)=c("setosa","versicolor","virginica")
groups

summary(data[data$Spec=="setosa",])

#Merging data frames along columns
df1=read.csv("iris - Col1.CSV")
head(df1)
```

```r
df2=read.csv("iris - Col2.CSV")
head(df2)

dfm1=merge(df1,df2,by="Id")
head(dfm1)

df3=read.csv("iris - Col3.CSV")
head(df3)

head(df1)

dfm2=merge(df1,df3,by=c("Id","Species"))
head(dfm2)

df4=read.csv("iris - Col4.CSV")
head(df4)

head(df1)

dfm3=cbind(df1,df4)
head(dfm3)

#Merging data frames along rows
df1=read.csv("iris - Row1.CSV")
head(df1)

df2=read.csv("iris - Row2.CSV")
head(df2)

rbind(df1,df2)

#Unique values & counts in a categorical column
head(data)

unique(data$Spec)

table(data$Spec)

#Getting samples of data frames
dim(data)

index=1:150
id_sample=sample(index,round(0.5*length(index)))
id_sample

data_sam1=data[id_sample,]
data_sam2=data[-id_sample,]

dim(data_sam1)

dim(data_sam2)
```

```r
#Write to a new CSV file
head(data)

write.csv(data,"iris_new.CSV",)

#Importing text files
data1=read.table("Data01.txt",sep=",",header = TRUE)
data1

data1=read.table("Data02.txt",sep=",")
data1

colnames(data1)=c("Col01","Col02","Col03")
data1
```