# Missing Data

## 1. Finding Missing Values

## Example   
# ---  
# Lets create a dataset dt  
# ---  
# OUR CODE GOES BELOW  
#  
Name <- c("John", "Tim", NA)  
Sex <- c("men", "men", "women")  
Age <- c(45, 53, NA)  
dt <- data.frame(Name, Sex, Age)  
  
# Then print out this dataset below  
dt

## Name Sex Age  
## 1 John men 45  
## 2 Tim men 53  
## 3 <NA> women NA

# Lets Identify missing data in your dataset   
# by using the function is.na()   
# ---  
#   
is.na(dt)

## Name Sex Age  
## [1,] FALSE FALSE FALSE  
## [2,] FALSE FALSE FALSE  
## [3,] TRUE FALSE TRUE

# Example   
# ---  
# We can also find out total missing values in each column   
# by using the function colSums()  
# ---  
# OUR CODE GOES BELOW  
#  
colSums(is.na(dt))

## Name Sex Age   
## 1 0 1

## 2. Dealing with Missing Values

## Example   
# ---  
# Question: Show all rows from the dataset which don't contain any missing values   
# ---  
#  
na.omit(dt)

## Name Sex Age  
## 1 John men 45  
## 2 Tim men 53

## Example   
# ---  
# Question: Recode/fill the missing value in a column with a number  
# ---  
# OUR CODE GOES BELOW  
#  
dt$Age[is.na(dt$Age)] <- 99  
  
dt

## Name Sex Age  
## 1 John men 45  
## 2 Tim men 53  
## 3 <NA> women 99

## Example  
# ---  
# Question: Recode or fill the missing value in a column with the mean value of the column-#-  
# ---  
# fill with mean  
#  
dt$Age[is.na(dt$Age)] <- mean(dt$Age, na.rm = TRUE)  
  
# print the dt table below  
dt

## Name Sex Age  
## 1 John men 45  
## 2 Tim men 53  
## 3 <NA> women 99

## Challenge 1  
# ---  
# Question: Using the given bus dataset below, recode the missing values of the payment\_method   
# and travel\_to columns with athen appropriate values  
# ---  
# OUR CODE GOES BELOW  
#  
  
# Lets first of all import our data table  
# ---  
#  
library("data.table")  
bus\_dataset <-fread('http://bit.ly/BusNairobiWesternTransport')  
  
# First check have a look at the dataset  
# --  
#  
head(bus\_dataset)

## ride\_id seat\_number payment\_method payment\_receipt travel\_date travel\_time  
## 1: 1442 15A Mpesa UZUEHCBUSO 0017-10-17 7:15  
## 2: 5437 14A Mpesa TIHLBUSGTE 0019-11-17 7:12  
## 3: 5710 8B Mpesa EQX8Q5G19O 0026-11-17 7:05  
## 4: 5777 19A Mpesa SGP18CL0ME 0027-11-17 7:10  
## 5: 5778 11A Mpesa BM97HFRGL9 0027-11-17 7:12  
## 6: 5777 18B Mpesa B6PBDU30IZ 0027-11-17 7:10  
## travel\_from travel\_to car\_type max\_capacity  
## 1: Migori Nairobi Bus 49  
## 2: Migori Nairobi Bus 49  
## 3: Keroka Nairobi Bus 49  
## 4: Homa Bay Nairobi Bus 49  
## 5: Migori Nairobi Bus 49  
## 6: Homa Bay Nairobi Bus 49

colSums(is.na(bus\_dataset))

## ride\_id seat\_number payment\_method payment\_receipt travel\_date   
## 0 0 0 0 0   
## travel\_time travel\_from travel\_to car\_type max\_capacity   
## 0 0 0 0 0

## Challenge 2  
# ---  
# Question: Clean the given dataset   
# ---  
# Dataset url = http://bit.ly/MS-PropertyDataset  
ms <- fread("http://bit.ly/MS-PropertyDataset")  
ms

## PID ST\_NUM ST\_NAME OWN\_OCCUPIED NUM\_BEDROOMS NUM\_BATH SQ\_FT  
## 1: 100001000 104 PUTNAM Y 3 1 1000  
## 2: 100002000 197 LEXINGTON N 3 1.5 --  
## 3: 100003000 NA LEXINGTON N n/a 1 850  
## 4: 100004000 201 BERKELEY 12 1 NaN 700  
## 5: NA 203 BERKELEY Y 3 2 1600  
## 6: 100006000 207 BERKELEY Y <NA> 1 800  
## 7: 100007000 NA WASHINGTON 2 HURLEY 950  
## 8: 100008000 213 TREMONT Y 1 1   
## 9: 100009000 215 TREMONT Y na 2 1800

colSums(is.na(ms))

## PID ST\_NUM ST\_NAME OWN\_OCCUPIED NUM\_BEDROOMS NUM\_BATH   
## 1 2 0 0 1 0   
## SQ\_FT   
## 0

#impute pid with 100005000  
ms$PID[is.na(ms$PID)] <- 100005000  
  
ms

## PID ST\_NUM ST\_NAME OWN\_OCCUPIED NUM\_BEDROOMS NUM\_BATH SQ\_FT  
## 1: 100001000 104 PUTNAM Y 3 1 1000  
## 2: 100002000 197 LEXINGTON N 3 1.5 --  
## 3: 100003000 NA LEXINGTON N n/a 1 850  
## 4: 100004000 201 BERKELEY 12 1 NaN 700  
## 5: 100005000 203 BERKELEY Y 3 2 1600  
## 6: 100006000 207 BERKELEY Y <NA> 1 800  
## 7: 100007000 NA WASHINGTON 2 HURLEY 950  
## 8: 100008000 213 TREMONT Y 1 1   
## 9: 100009000 215 TREMONT Y na 2 1800

ms$ST\_NUM[is.na(ms$ST\_NUM)] <- 197  
ms

## PID ST\_NUM ST\_NAME OWN\_OCCUPIED NUM\_BEDROOMS NUM\_BATH SQ\_FT  
## 1: 100001000 104 PUTNAM Y 3 1 1000  
## 2: 100002000 197 LEXINGTON N 3 1.5 --  
## 3: 100003000 197 LEXINGTON N n/a 1 850  
## 4: 100004000 201 BERKELEY 12 1 NaN 700  
## 5: 100005000 203 BERKELEY Y 3 2 1600  
## 6: 100006000 207 BERKELEY Y <NA> 1 800  
## 7: 100007000 197 WASHINGTON 2 HURLEY 950  
## 8: 100008000 213 TREMONT Y 1 1   
## 9: 100009000 215 TREMONT Y na 2 1800

ms$NUM\_BEDROOMS[is.na(ms$NUM\_BEDROOMS)] <- mean(ms$NUM\_BEDROOMS, na.rm = TRUE)

## Warning in mean.default(ms$NUM\_BEDROOMS, na.rm = TRUE): argument is not numeric  
## or logical: returning NA

ms

## PID ST\_NUM ST\_NAME OWN\_OCCUPIED NUM\_BEDROOMS NUM\_BATH SQ\_FT  
## 1: 100001000 104 PUTNAM Y 3 1 1000  
## 2: 100002000 197 LEXINGTON N 3 1.5 --  
## 3: 100003000 197 LEXINGTON N n/a 1 850  
## 4: 100004000 201 BERKELEY 12 1 NaN 700  
## 5: 100005000 203 BERKELEY Y 3 2 1600  
## 6: 100006000 207 BERKELEY Y <NA> 1 800  
## 7: 100007000 197 WASHINGTON 2 HURLEY 950  
## 8: 100008000 213 TREMONT Y 1 1   
## 9: 100009000 215 TREMONT Y na 2 1800

## Challenge 3  
# ---  
# Question:   
# ---  
# Dataset url = http://bit.ly/AirQualityDataset  
  
air <- fread("http://bit.ly/AirQualityDataset")  
air

## Ozone Solar.R Wind Temp Month Day  
## 1: 41 190 7.4 67 5 1  
## 2: 36 118 8.0 72 5 2  
## 3: 12 149 12.6 74 5 3  
## 4: 18 313 11.5 62 5 4  
## 5: NA NA 14.3 56 5 5  
## ---   
## 149: 30 193 6.9 70 9 26  
## 150: NA 145 13.2 77 9 27  
## 151: 14 191 14.3 75 9 28  
## 152: 18 131 8.0 76 9 29  
## 153: 20 223 11.5 68 9 30

colSums(is.na(air))

## Ozone Solar.R Wind Temp Month Day   
## 37 7 0 0 0 0

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

air$Solar.R[is.na(air$Solar.R)] <- as.integer( mean(air$Solar.R, na.rm = TRUE))  
  
air

## Ozone Solar.R Wind Temp Month Day  
## 1: 41 190 7.4 67 5 1  
## 2: 36 118 8.0 72 5 2  
## 3: 12 149 12.6 74 5 3  
## 4: 18 313 11.5 62 5 4  
## 5: 42 185 14.3 56 5 5  
## ---   
## 149: 30 193 6.9 70 9 26  
## 150: 42 145 13.2 77 9 27  
## 151: 14 191 14.3 75 9 28  
## 152: 18 131 8.0 76 9 29  
## 153: 20 223 11.5 68 9 30

colSums(is.na(air))

## Ozone Solar.R Wind Temp Month Day   
## 0 0 0 0 0 0