# NA vignette

## Contents



Most of this code came from Harvard STAT 109 class, Prof. Bharatendra Rai. Material used here for educational purposes. It is available in YouTube and GitHub. See links under references. I expanded the material with my own notes and R documentation and I plan to continue adding examples overtime.

# 1 Datasets from DAAG to test NA

```
data(package='DAAG')
```

## 1.1 Rainforest example

### 1.1.1 Looking for NA and missing data

Reference: Dr. Bharantendra https://www.youtube.com/watch?v=Q7gYkpSi8Nk&list=PL34t5iLfZddtUUABMikey6NtL05h index=11

```
data('rainforest')
# str(rainforest)
summary(rainforest)
```

```
##       dbh             wood             bark             root
##  Min.   : 4.00   Min.   :   3.0   Min.   :  8.00   Min.   :  2.00
##  1st Qu.: 8.00   1st Qu.:  29.0   1st Qu.: 11.75   1st Qu.:  6.00
##  Median :12.00   Median : 100.0   Median : 45.50   Median : 16.00
##  Mean   :16.06   Mean   : 265.4   Mean   : 51.00   Mean   : 30.85
##  3rd Qu.:22.00   3rd Qu.: 386.5   3rd Qu.: 84.75   3rd Qu.: 44.00
##  Max.   :56.00   Max.   :1530.0   Max.   :105.00   Max.   :135.00
##                  NA's   :1        NA's   :61       NA's   :52
##      rootsk           branch                species
```

```
##  Min.   : 0.300   Min.   :  4.00   Acacia mabellae:16
##  1st Qu.: 1.300   1st Qu.:  9.00   C. fraseri     :12
##  Median : 2.400   Median : 25.00   Acmena smithii :26
##  Mean   : 7.477   Mean   : 32.86   B. myrtifolia  :11
##  3rd Qu.:13.000   3rd Qu.: 45.50
##  Max.   :24.000   Max.   :120.00
##  NA's   :52       NA's   :22
```

```
dim(rainforest)
```

```
## [1] 65  7
```

Notice the extra row with NA values. You will not be able to calculate the mean or the sd.

### 1.1.2 Ignoring NA

```
mean(rainforest$wood, na.rm = TRUE)
```

```
## [1] 265.3906
```

## 1.2 Science example

```
# Load the data
data('science')
summary(science)
```

```
##  State         PrivPub        school    class     sex           like
##  ACT:1336   private:452   36     :123   1:881   f   :691   Min.   :1.000
##  NSW:  49   public :933   16     : 94   2:347   m   :692   1st Qu.:4.000
##                           23     : 87   3: 88   NA's:  2   Median :5.000
##                           17     : 50   4: 69              Mean   :5.082
##                           31     : 50                      3rd Qu.:6.000
##                           3      : 49                      Max.   :8.000
##                           (Other):932
##      Class
##  17.1   :  50
##  36.2   :  44
##  21.1   :  42
##  36.1   :  37
##  32.1   :  34
##  25.1   :  31
##  (Other):1147
```

```
dim(science)
```

```
## [1] 1385    7
```

### 1.2.1 Identify rows with NA

Use !complete.cases(dataset) which mean NOT complete cases, those with NA.

```
science[!complete.cases(science),]
```

```
##     State PrivPub school class  sex like Class
## 671   ACT  public     19     1 <NA>    5  19.1
## 672   ACT  public     19     1 <NA>    5  19.1
```

### 1.2.2 Save dataset without NA

```
science_wo_na <- na.omit(science)
summary(science_wo_na)
```

```
##  State         PrivPub        school    class   sex          like
##  ACT:1334   private:452   36     :123   1:879   f:691   Min.   :1.000
##  NSW:  49   public :931   16     : 94   2:347   m:692   1st Qu.:4.000
##                           23     : 87   3: 88           Median :5.000
##                           17     : 50   4: 69           Mean   :5.082
##                           31     : 50                   3rd Qu.:6.000
##                           3      : 49                   Max.   :8.000
##                           (Other):930
##      Class
##  17.1   :  50
##  36.2   :  44
##  21.1   :  42
##  36.1   :  37
##  32.1   :  34
##  25.1   :  31
##  (Other):1145
```

```
dim(science_wo_na)
```

```
## [1] 1383    7
```