DATA SCIENCE WITH R

Class 4 – Data Manipulation in R

Topic 5

★ Missing Value Treatment ★

INDEX

Manipulating data using base R

Using dplyr to manipulate data

Working with date objects

Merging tables

➔ **Missing value treatment**

Using reshape2() to transpose data

Manipulating Character Strings

Using sqldf

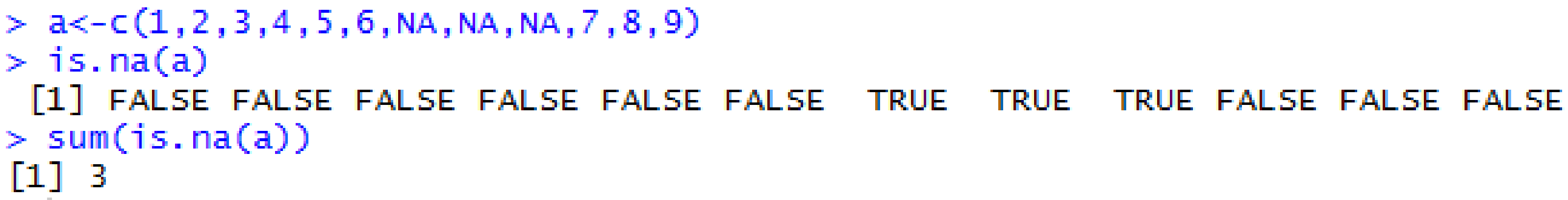
Missing Values

Missing Values

- Identifying missing values in a column
- Imputing missing values

Missing Values

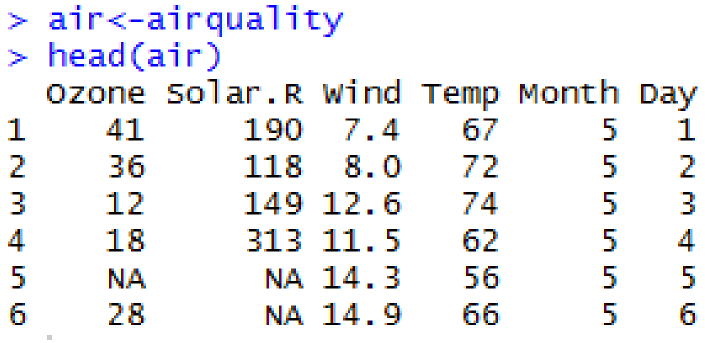
- Using is.na() to find out the total number of missing values



```
> a<-c(1,2,3,4,5,6,NA,NA,NA,7,8,9)
> is.na(a)
 [1] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE
> sum(is.na(a))
[1] 3
```

Missing Values

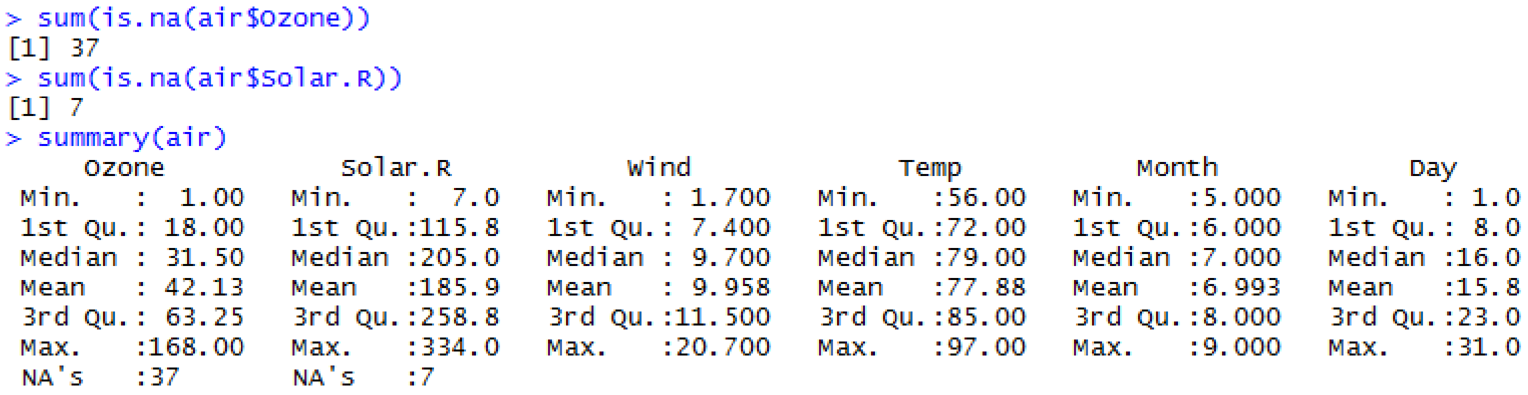
- Using is.na() to find out the total number of missing values



```
> air<-airquality
> head(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
6    28      NA 14.9   66     5   6
```

Missing Values

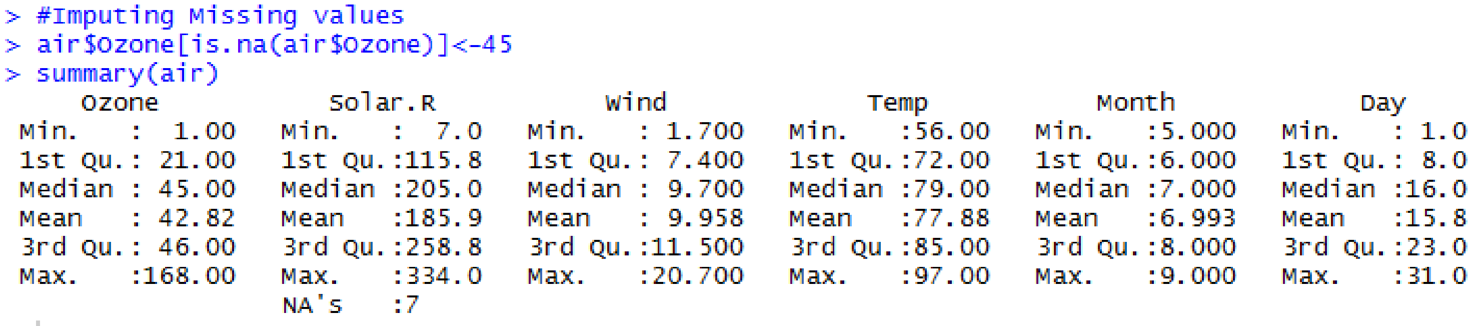
- Using is.na() to find out the total number of missing values



```
> sum(is.na(air$Ozone))
[1] 37
> sum(is.na(air$Solar.R))
[1] 7
> summary(air)
     Ozone            Solar.R           Wind             Temp           Month            Day      
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00   Min.   :5.000   Min.   : 1.0  
 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00   1st Qu.:6.000   1st Qu.: 8.0  
 Median : 31.50   Median :205.0   Median : 9.700   Median :79.00   Median :7.000   Median :16.0  
 Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88   Mean   :6.993   Mean   :15.8  
 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00   3rd Qu.:8.000   3rd Qu.:23.0  
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00   Max.   :9.000   Max.   :31.0  
 NA's   :37       NA's   :7                                                                       
```

Missing Values

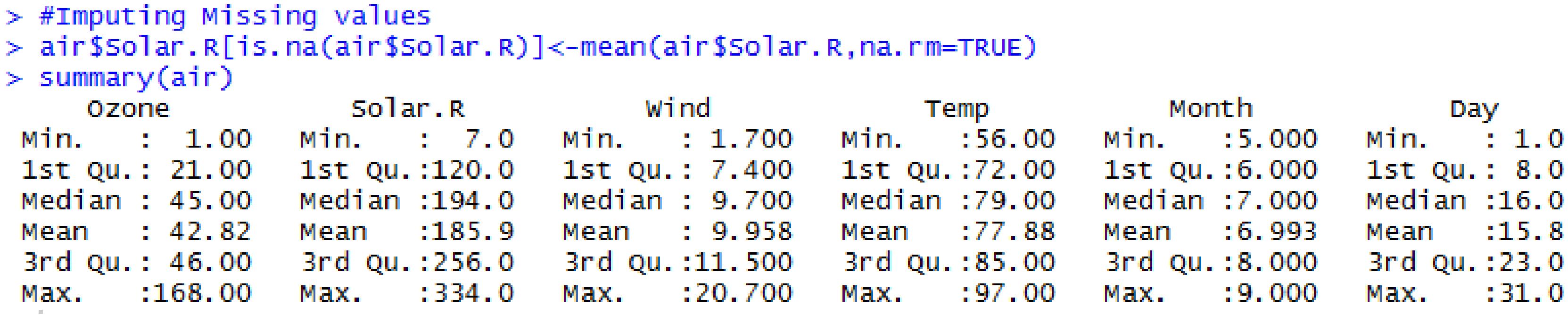
- Imputing missing values



```
> #Imputing Missing values
> air$Ozone[is.na(air$Ozone)]<-45
> summary(air)
     Ozone           Solar.R           Wind             Temp           Month            Day      
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00   Min.   :5.000   Min.   : 1.0  
 1st Qu.: 21.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00   1st Qu.:6.000   1st Qu.: 8.0  
 Median : 45.00   Median :205.0   Median : 9.700   Median :79.00   Median :7.000   Median :16.0  
 Mean   : 42.82   Mean   :185.9   Mean   : 9.958   Mean   :77.88   Mean   :6.993   Mean   :15.8  
 3rd Qu.: 46.00   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00   3rd Qu.:8.000   3rd Qu.:23.0  
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00   Max.   :9.000   Max.   :31.0  
                  NA's   :7                                                                      
```

Missing Values

- Imputing missing values



```
> #Imputing Missing values
> air$Solar.R[is.na(air$Solar.R)]<-mean(air$Solar.R,na.rm=TRUE)
> summary(air)
     Ozone           Solar.R           Wind             Temp           Month            Day      
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00   Min.   :5.000   Min.   : 1.0  
 1st Qu.: 21.00   1st Qu.:120.0   1st Qu.: 7.400   1st Qu.:72.00   1st Qu.:6.000   1st Qu.: 8.0  
 Median : 45.00   Median :194.0   Median : 9.700   Median :79.00   Median :7.000   Median :16.0  
 Mean   : 42.82   Mean   :185.9   Mean   : 9.958   Mean   :77.88   Mean   :6.993   Mean   :15.8  
 3rd Qu.: 46.00   3rd Qu.:256.0   3rd Qu.:11.500   3rd Qu.:85.00   3rd Qu.:8.000   3rd Qu.:23.0  
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00   Max.   :9.000   Max.   :31.0  
```

DATA
SCIENCE
WITH R

RECAP

- Identifying missing values in a column
- Imputing missing values