15. Missing data imputation

Missing value imputation

```
math=c(88,95,85,NA,76,69,NA,70,68)
mean(math)
```

```
[1] NA
```

```
median(math)
```

```
[1] NA
```

```
sd(math)
```

```
[1] NA
```

```
math[is.na(math)]
```

```
[1] NA NA
```

```
math[!is.na(math)]
```

```
[1] 88 95 85 76 69 70 68
```

```
mean(math,na.rm=T)
```

```
[1] 78.71429
```

```
#median(math,na.rm=T)
#sd(math,na.rm=T)
```

Ex1. Missing value imputation

```
math1=math
math1[is.na(math1)]=mean(math,na.rm=T)
math1
```

```
[1] 88.00000 95.00000 85.00000 78.71429 76.00000 69.00000 78.71429 70.00000
[9] 68.00000
```

```
#math2=math
#math2[is.na(math2)]=median(math,na.rm=T)
#math2
```

Ex2. Missing value imputation

```
var1=c(19,13,NA,17,5,16,NA,20,18,12,25,12,30,22)
var2=c(49,53,50,48,NA,51,53,51,47,NA,44,50,52,NA)
data=data.frame(var1,var2)
summary(data)
```

```
      var1            var2
 Min.   : 5.00   Min.   :44.00
 1st Qu.:12.75   1st Qu.:48.50
 Median :17.50   Median :50.00
 Mean   :17.42   Mean   :49.82
 3rd Qu.:20.50   3rd Qu.:51.50
 Max.   :30.00   Max.   :53.00
 NA's   :2       NA's   :3
```

```
data$var1[is.na(data$var1)]=mean(data$var1[!is.na(data$var1)])
data$var2[is.na(data$var2)]=median(data$var2[!is.na(data$var2)])
data
```

```
       var1 var2
1  19.00000   49
2  13.00000   53
3  17.41667   50
4  17.00000   48
5   5.00000   50
6  16.00000   51
7  17.41667   53
8  20.00000   51
9  18.00000   47
10 12.00000   50
11 25.00000   44
12 12.00000   50
13 30.00000   52
14 22.00000   50
```

Ex3. Filling all missing values - Dataset car missing

```
car_missing=read.csv(file.choose())
head(car_missing,n=10)
```

```
   Manufact   Model  Sales Resales
1         1 Integra 16.919  16.360
2         1      TL 39.384  19.875
```

```
3         1      CL 14.114  18.225
4         1      RL  8.588  29.725
5         2      A4 20.397  22.255
6         2      A6 18.780  23.555
7         2      A8  1.380  39.000
8         3    323i 19.747      NA
9         3    328i  9.231  28.675
10        3    528i 17.527  36.125
```

```
tail(car_missing,n=10)
```

```
    Manufact  Model  Sales Resales
148       29 Passat 51.102  16.725
149       29 Cabrio  9.569  16.575
150       29    GTI  5.596  13.760
151       29 Beetle 49.463      NA
152       30    S40 16.957      NA
153       30    V40  3.545      NA
154       30    S70 15.245      NA
155       30    V70 17.531      NA
156       30    C70  3.493      NA
157       30    S80 18.969      NA
```

```
summary(car_missing)
```

```
    Manufact        Model               Sales           Resales
 Min.   : 1.00   Length:157         Min.   :  0.11   Min.   : 5.16
 1st Qu.: 8.00   Class :character   1st Qu.: 14.11   1st Qu.:11.26
 Median :16.00   Mode  :character   Median : 29.45   Median :14.18
 Mean   :15.54                      Mean   : 53.00   Mean   :18.07
 3rd Qu.:22.00                      3rd Qu.: 67.96   3rd Qu.:19.88
 Max.   :30.00                      Max.   :540.56   Max.   :67.55
                                                     NA's   :36
```

```
allmn=mean(car_missing$Resales,na.rm = T)
allmn
```

```
[1] 18.07298
```

```
allmis=which(is.na(car_missing$Resales))
allmis
```

```
 [1]   8  19  28  35  45  51  67  73  75  76  79  97  98  99 100 101 107 108 110
[20] 111 118 124 128 129 133 134 135 136 142 151 152 153 154 155 156 157
```

```
for (i in 1:length(allmis)) car_missing$Resales[allmis[i]]=allmn
head(car_missing,n=10)
```

```
   Manufact   Model  Sales  Resales
1         1 Integra 16.919 16.36000
```

```
2         1      TL 39.384 19.87500
3         1      CL 14.114 18.22500
4         1      RL  8.588 29.72500
5         2      A4 20.397 22.25500
6         2      A6 18.780 23.55500
7         2      A8  1.380 39.00000
8         3    323i 19.747 18.07298
9         3    328i  9.231 28.67500
10        3    528i 17.527 36.12500
```

```
tail(car_missing,n=10)
```

```
    Manufact  Model  Sales  Resales
148       29 Passat 51.102 16.72500
149       29 Cabrio  9.569 16.57500
150       29    GTI  5.596 13.76000
151       29 Beetle 49.463 18.07298
152       30    S40 16.957 18.07298
153       30    V40  3.545 18.07298
154       30    S70 15.245 18.07298
155       30    V70 17.531 18.07298
156       30    C70  3.493 18.07298
157       30    S80 18.969 18.07298
```

```
write.csv(x=car_missing,file="cars.csv")
```

Ex4. Filling particular missing values - Dataset car missing

```
car_missing=read.csv(file.choose())
head(car_missing,n=10)
```

```
   Manufact   Model  Sales Resales
1         1 Integra 16.919  16.360
2         1      TL 39.384  19.875
3         1      CL 14.114  18.225
4         1      RL  8.588  29.725
5         2      A4 20.397  22.255
6         2      A6 18.780  23.555
7         2      A8  1.380  39.000
8         3    323i 19.747      NA
9         3    328i  9.231  28.675
10        3    528i 17.527  36.125
```

```
tail(car_missing,n=10)
```

```
    Manufact  Model  Sales Resales
148       29 Passat 51.102  16.725
149       29 Cabrio  9.569  16.575
150       29    GTI  5.596  13.760
```

```
151       29 Beetle 49.463      NA
152       30    S40 16.957      NA
153       30    V40  3.545      NA
154       30    S70 15.245      NA
155       30    V70 17.531      NA
156       30    C70  3.493      NA
157       30    S80 18.969      NA
```

```
summary(car_missing)
```

```
    Manufact        Model              Sales           Resales
 Min.   : 1.00   Length:157         Min.   :  0.11   Min.   : 5.16
 1st Qu.: 8.00   Class :character   1st Qu.: 14.11   1st Qu.:11.26
 Median :16.00   Mode  :character   Median : 29.45   Median :14.18
 Mean   :15.54                      Mean   : 53.00   Mean   :18.07
 3rd Qu.:22.00                      3rd Qu.: 67.96   3rd Qu.:19.88
 Max.   :30.00                      Max.   :540.56   Max.   :67.55
                                                     NA's   :36
```

```
mnmis3=subset(car_missing,Manufact==3,select = c(Resales))
mnmis3
```

```
   Resales
8       NA
9   28.675
10  36.125
```

```
mnmis3=sapply(mnmis3,FUN=mean,na.rm=T)
mnmis3
```

```
Resales
   32.4
```

```
#mnmis3=sapply(subset(car_missing,Manufact==3,select = c(Resales)),FUN=mean,na.rm=T)
mis=which(car_missing$Manufact==3 & is.na(car_missing$Resales))
car_missing$Resales[mis]=mnmis3
head(car_missing,n=10)
```

```
   Manufact   Model  Sales Resales
1         1 Integra 16.919  16.360
2         1      TL 39.384  19.875
3         1      CL 14.114  18.225
4         1      RL  8.588  29.725
5         2      A4 20.397  22.255
6         2      A6 18.780  23.555
7         2      A8  1.380  39.000
8         3    323i 19.747  32.400
9         3    328i  9.231  28.675
10        3    528i 17.527  36.125
```