Data Cleaning

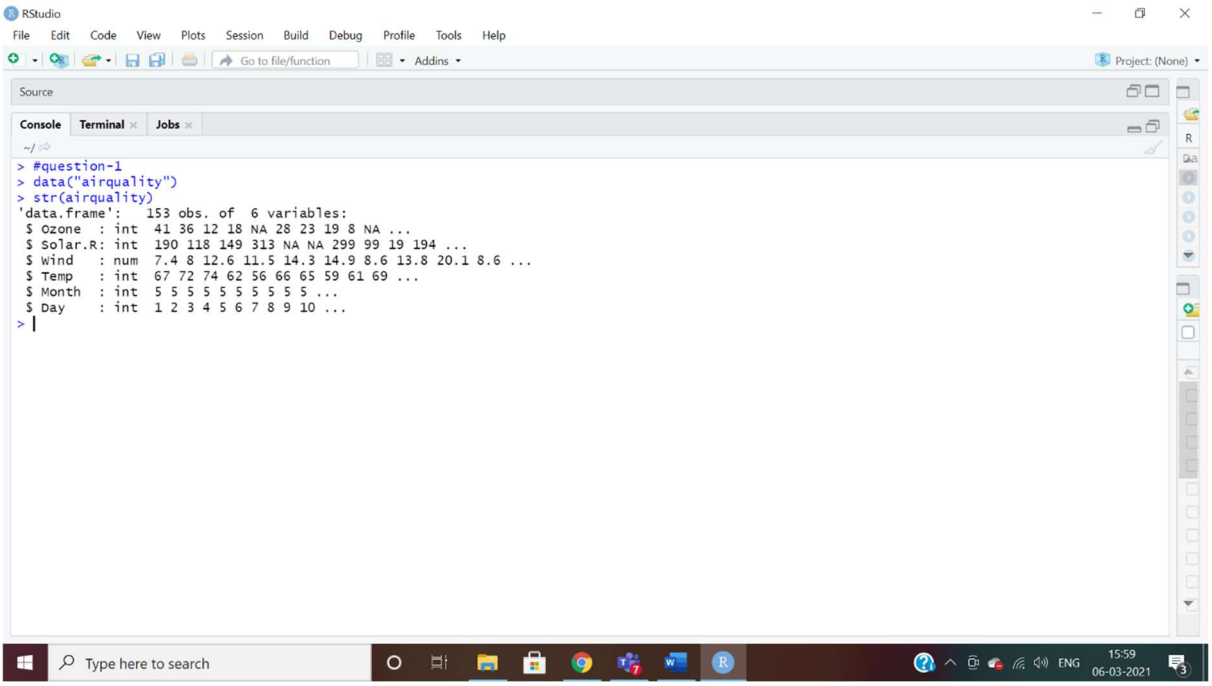
In this study, we will use airquality dataset.

· Print the strcture of the dataset

Answer: data("airquality")

str(airquality)

OUTPUT



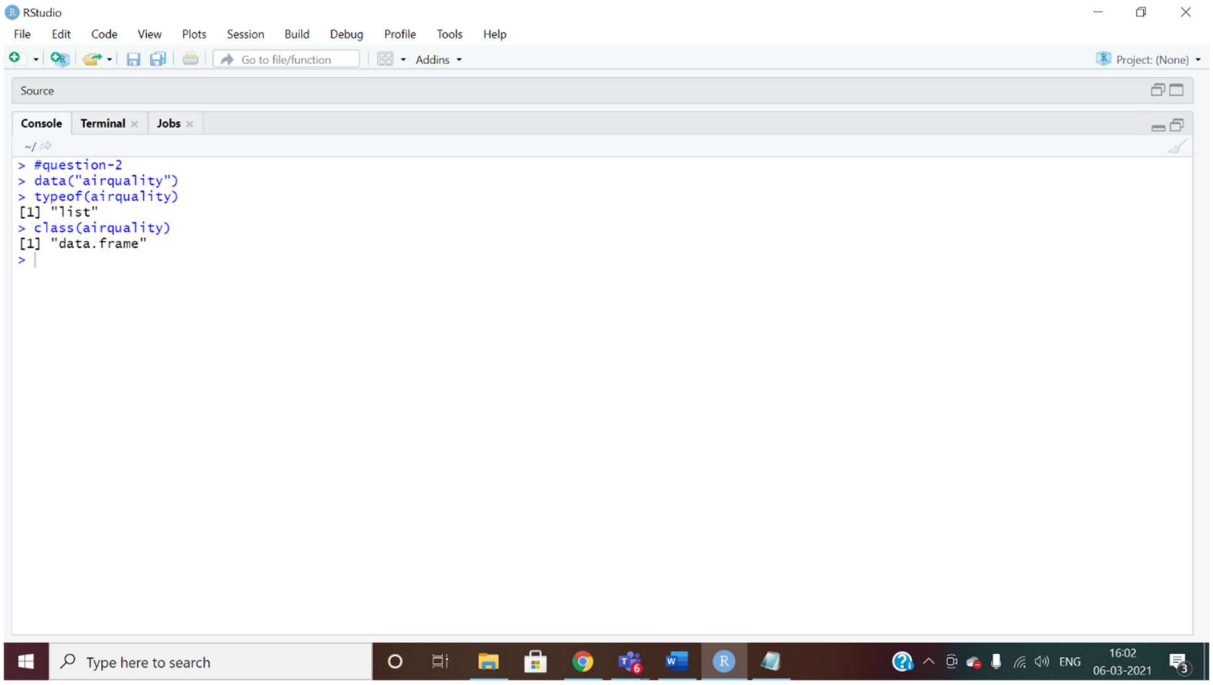
· What is the datatype of the datset?

Answer: data("airquality")

typeof(airquality)

class(airquality)

OUTPUT



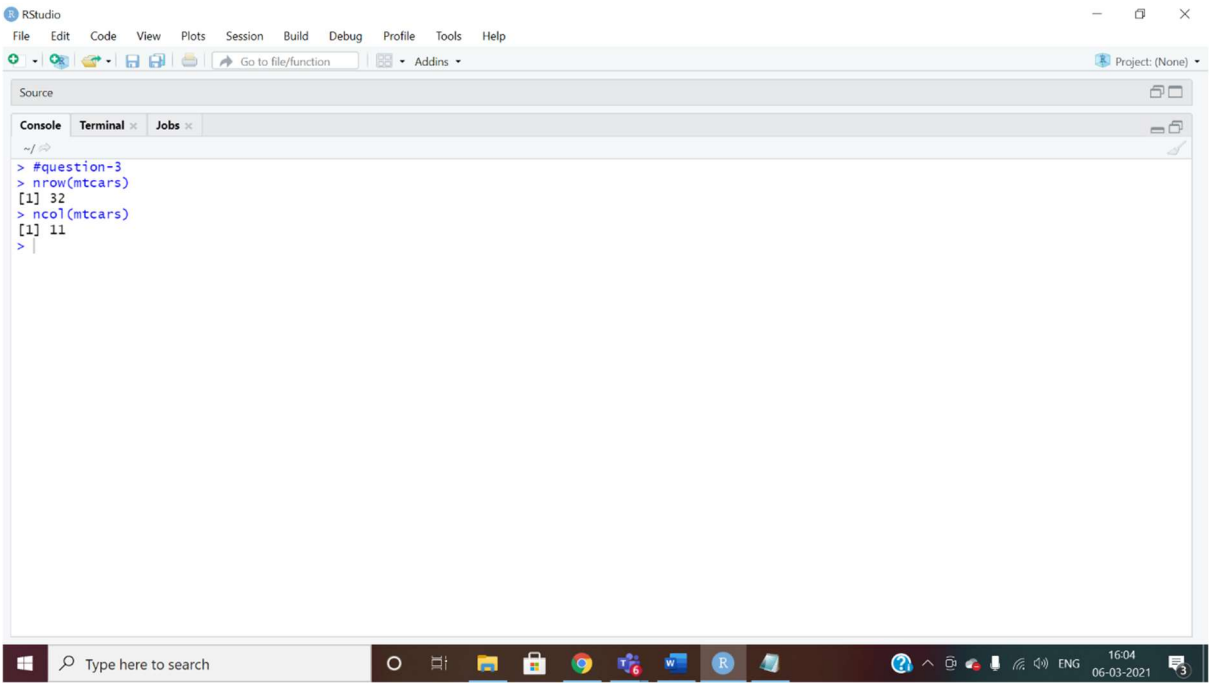
· How many colmns and rows are there in the dataset??

Answer:

nrow(airquality)

ncol(airquality)

OUTPUT

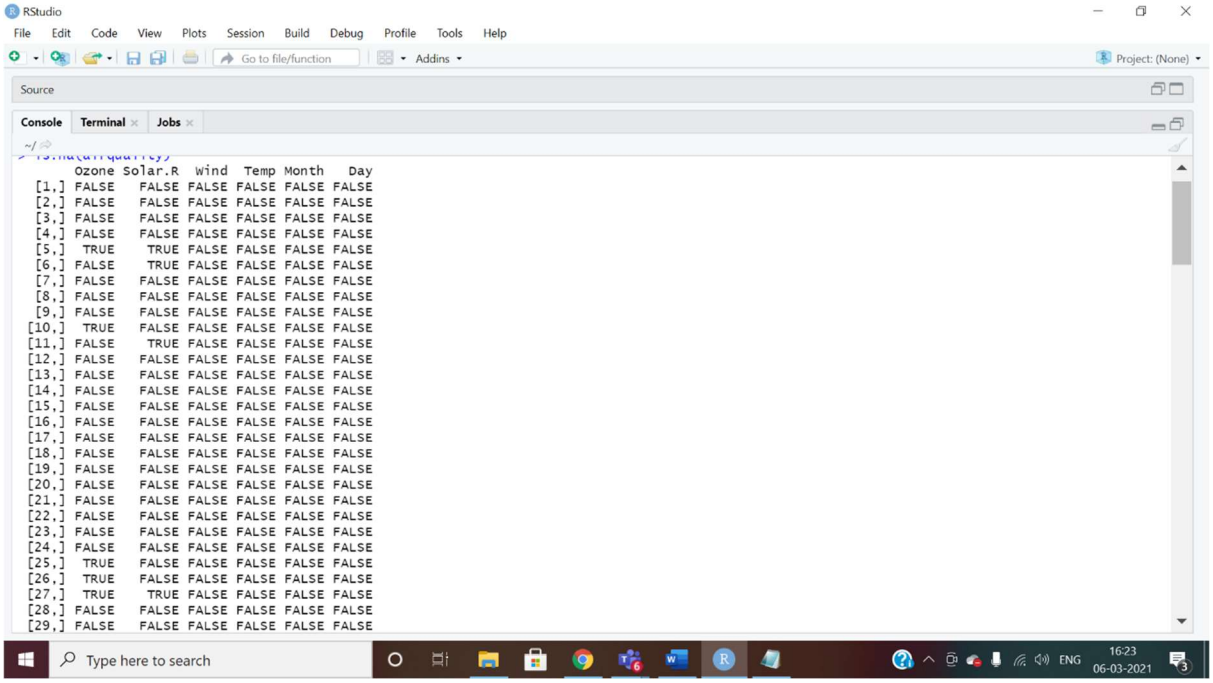


· **Use the function is.na() to find whether any missing values are in the dataset airquality**

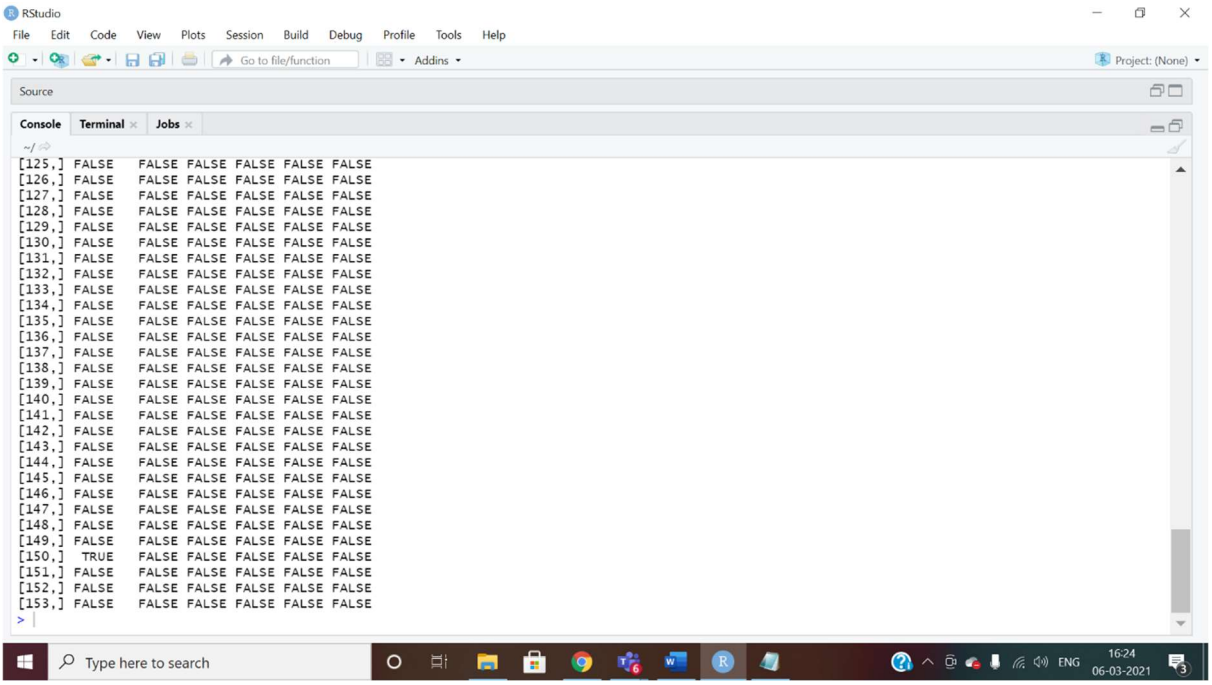
Answer:

is.na(airquality)

OUTPUT



```
      Ozone Solar.R  Wind  Temp Month   Day
  [1,] FALSE   FALSE FALSE FALSE FALSE FALSE
  [2,] FALSE   FALSE FALSE FALSE FALSE FALSE
  [3,] FALSE   FALSE FALSE FALSE FALSE FALSE
  [4,] FALSE   FALSE FALSE FALSE FALSE FALSE
  [5,]  TRUE    TRUE FALSE FALSE FALSE FALSE
  [6,] FALSE    TRUE FALSE FALSE FALSE FALSE
  [7,] FALSE   FALSE FALSE FALSE FALSE FALSE
  [8,] FALSE   FALSE FALSE FALSE FALSE FALSE
  [9,] FALSE   FALSE FALSE FALSE FALSE FALSE
 [10,]  TRUE   FALSE FALSE FALSE FALSE FALSE
 [11,] FALSE    TRUE FALSE FALSE FALSE FALSE
 [12,] FALSE   FALSE FALSE FALSE FALSE FALSE
 [13,] FALSE   FALSE FALSE FALSE FALSE FALSE
 [14,] FALSE   FALSE FALSE FALSE FALSE FALSE
 [15,] FALSE   FALSE FALSE FALSE FALSE FALSE
 [16,] FALSE   FALSE FALSE FALSE FALSE FALSE
 [17,] FALSE   FALSE FALSE FALSE FALSE FALSE
 [18,] FALSE   FALSE FALSE FALSE FALSE FALSE
 [19,] FALSE   FALSE FALSE FALSE FALSE FALSE
 [20,] FALSE   FALSE FALSE FALSE FALSE FALSE
 [21,] FALSE   FALSE FALSE FALSE FALSE FALSE
 [22,] FALSE   FALSE FALSE FALSE FALSE FALSE
 [23,] FALSE   FALSE FALSE FALSE FALSE FALSE
 [24,] FALSE   FALSE FALSE FALSE FALSE FALSE
 [25,]  TRUE   FALSE FALSE FALSE FALSE FALSE
 [26,]  TRUE   FALSE FALSE FALSE FALSE FALSE
 [27,]  TRUE    TRUE FALSE FALSE FALSE FALSE
 [28,] FALSE   FALSE FALSE FALSE FALSE FALSE
 [29,] FALSE   FALSE FALSE FALSE FALSE FALSE
```



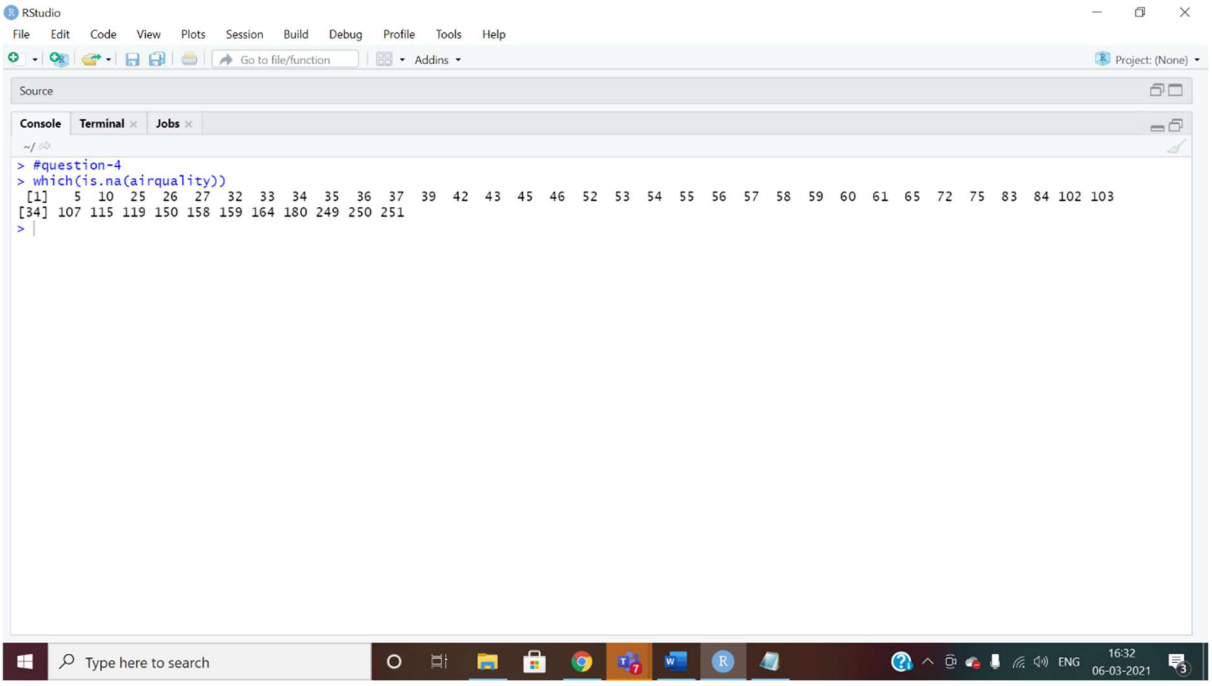
```
[125,] FALSE   FALSE FALSE FALSE FALSE FALSE
[126,] FALSE   FALSE FALSE FALSE FALSE FALSE
[127,] FALSE   FALSE FALSE FALSE FALSE FALSE
[128,] FALSE   FALSE FALSE FALSE FALSE FALSE
[129,] FALSE   FALSE FALSE FALSE FALSE FALSE
[130,] FALSE   FALSE FALSE FALSE FALSE FALSE
[131,] FALSE   FALSE FALSE FALSE FALSE FALSE
[132,] FALSE   FALSE FALSE FALSE FALSE FALSE
[133,] FALSE   FALSE FALSE FALSE FALSE FALSE
[134,] FALSE   FALSE FALSE FALSE FALSE FALSE
[135,] FALSE   FALSE FALSE FALSE FALSE FALSE
[136,] FALSE   FALSE FALSE FALSE FALSE FALSE
[137,] FALSE   FALSE FALSE FALSE FALSE FALSE
[138,] FALSE   FALSE FALSE FALSE FALSE FALSE
[139,] FALSE   FALSE FALSE FALSE FALSE FALSE
[140,] FALSE   FALSE FALSE FALSE FALSE FALSE
[141,] FALSE   FALSE FALSE FALSE FALSE FALSE
[142,] FALSE   FALSE FALSE FALSE FALSE FALSE
[143,] FALSE   FALSE FALSE FALSE FALSE FALSE
[144,] FALSE   FALSE FALSE FALSE FALSE FALSE
[145,] FALSE   FALSE FALSE FALSE FALSE FALSE
[146,] FALSE   FALSE FALSE FALSE FALSE FALSE
[147,] FALSE   FALSE FALSE FALSE FALSE FALSE
[148,] FALSE   FALSE FALSE FALSE FALSE FALSE
[149,] FALSE   FALSE FALSE FALSE FALSE FALSE
[150,]  TRUE   FALSE FALSE FALSE FALSE FALSE
[151,] FALSE   FALSE FALSE FALSE FALSE FALSE
[152,] FALSE   FALSE FALSE FALSE FALSE FALSE
[153,] FALSE   FALSE FALSE FALSE FALSE FALSE
>
```

· **Print the indices of the missing values in the dataset airquality in row major representation**

Answer

which(is.na(airquality))

OUTPUT



· **Print indices of the missing values in row and column number wise (Hint: Use function which() and argument arr.ind = TRUE)**

Answer:

which(is.na(airquality),TRUE)

OUTPUT

```
> #question-4
> which(is.na(airquality),TRUE)
      row col
 [1,]   5   1
 [2,]  10   1
 [3,]  25   1
 [4,]  26   1
 [5,]  27   1
 [6,]  32   1
 [7,]  33   1
 [8,]  34   1
 [9,]  35   1
[10,]  36   1
[11,]  37   1
[12,]  39   1
[13,]  42   1
[14,]  43   1
[15,]  45   1
[16,]  46   1
[17,]  52   1
[18,]  53   1
[19,]  54   1
[20,]  55   1
[21,]  56   1
[22,]  57   1
[23,]  58   1
[24,]  59   1
[25,]  60   1
[26,]  61   1
[27,]  65   1
```

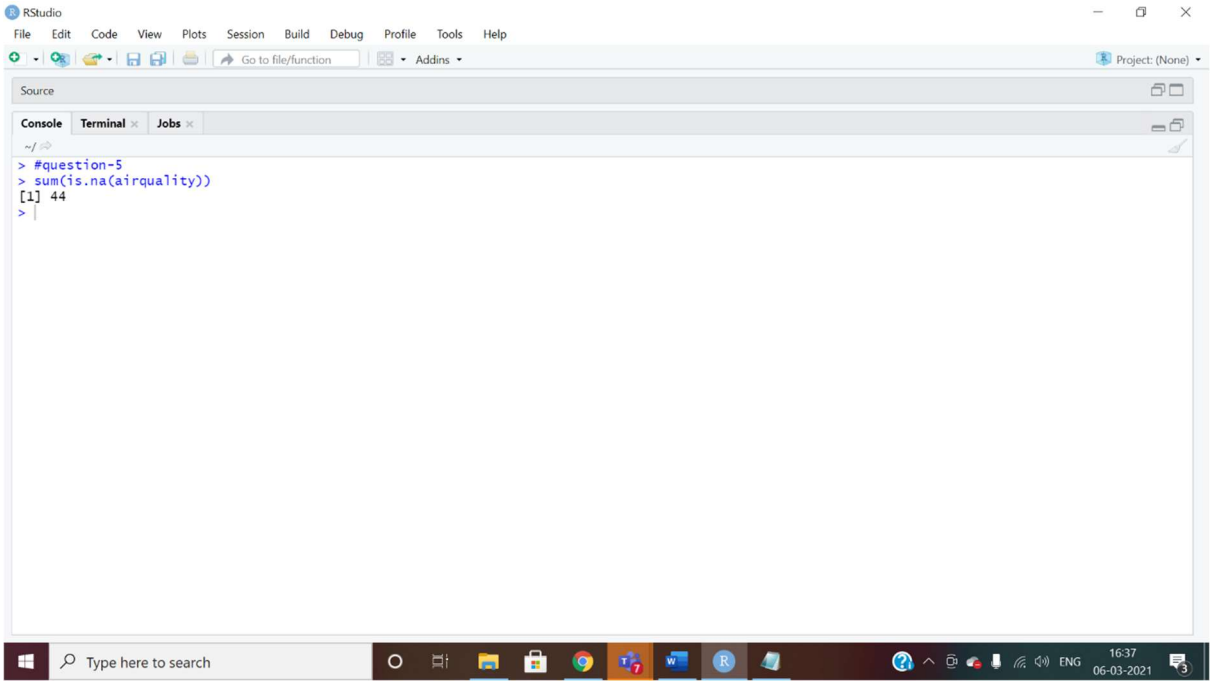
```
[16,]  46   1
[17,]  52   1
[18,]  53   1
[19,]  54   1
[20,]  55   1
[21,]  56   1
[22,]  57   1
[23,]  58   1
[24,]  59   1
[25,]  60   1
[26,]  61   1
[27,]  65   1
[28,]  72   1
[29,]  75   1
[30,]  83   1
[31,]  84   1
[32,] 102   1
[33,] 103   1
[34,] 107   1
[35,] 115   1
[36,] 119   1
[37,] 150   1
[38,]   5   2
[39,]   6   2
[40,]  11   2
[41,]  27   2
[42,]  96   2
[43,]  97   2
[44,]  98   2
>
```

· How many missing values are in the dataset airquality?

Answer:

sum(is.na(airquality))

OUTPUT

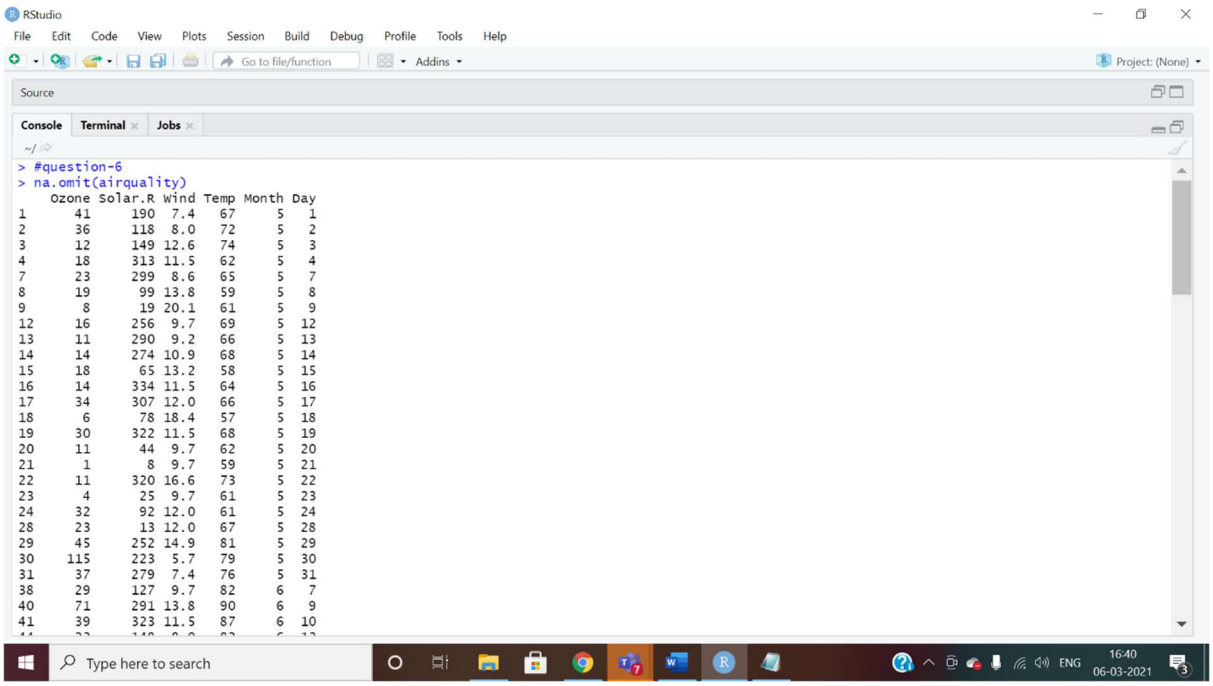


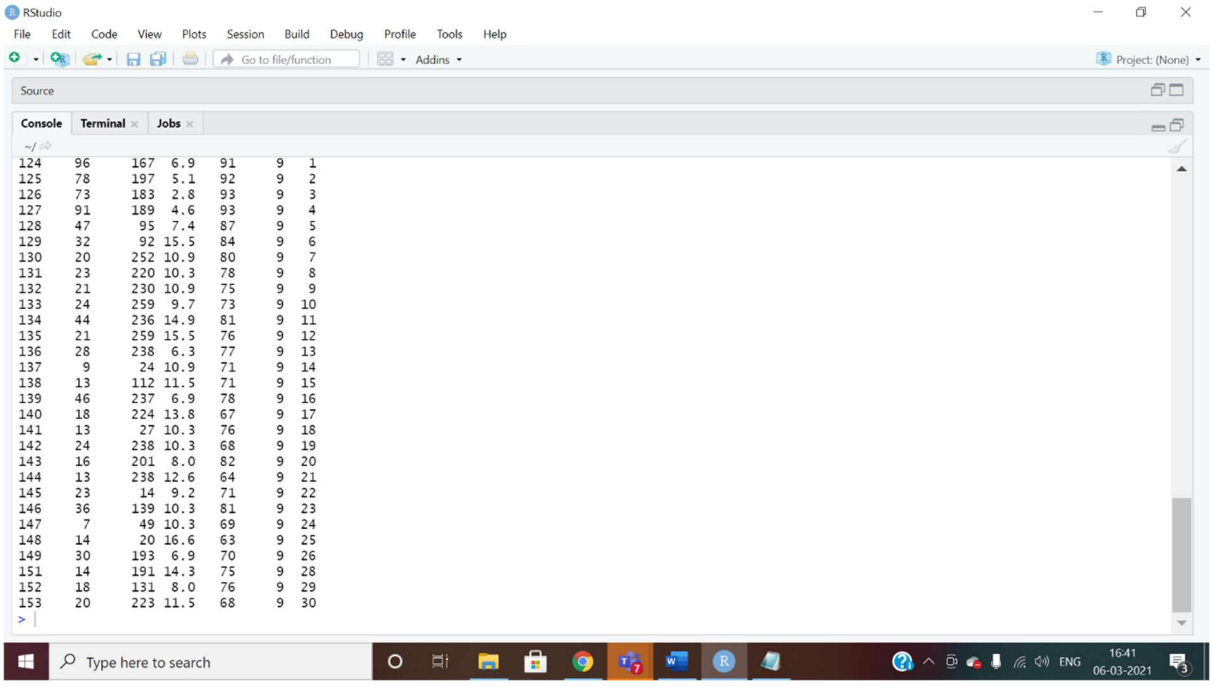
· How would you omit all rows containing missing values?

Answer:

na.omit(airquality)

OUTPUT





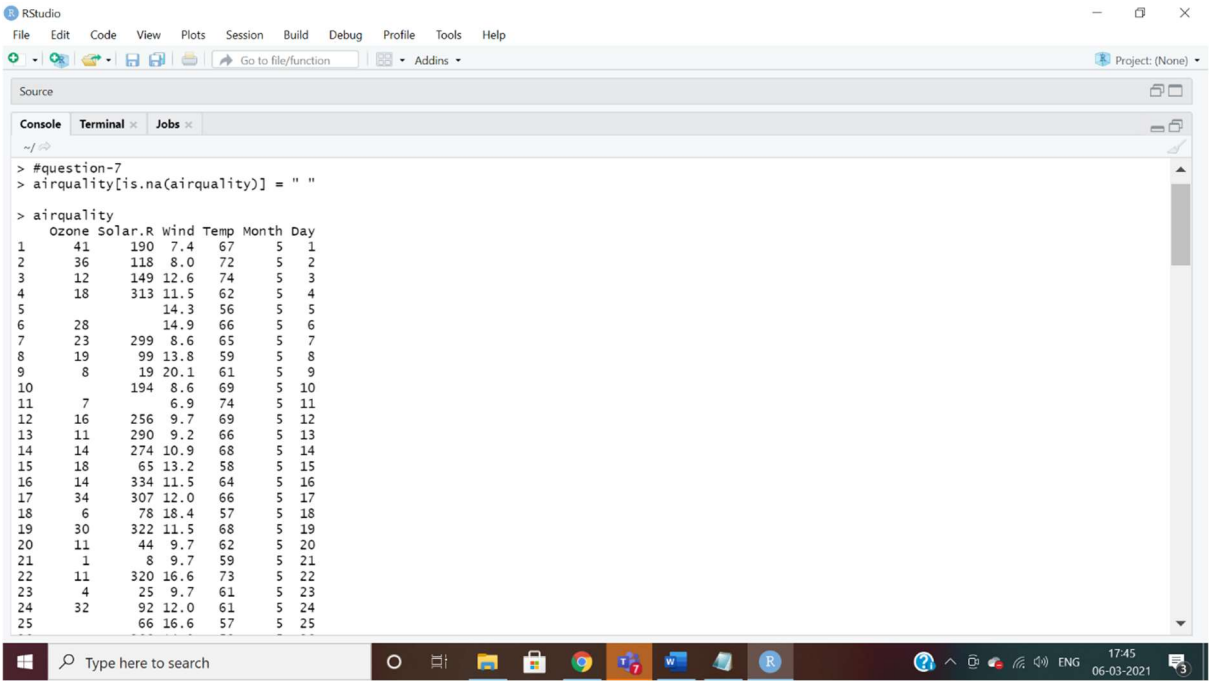
```
124    96     167  6.9   91     9   1
125    78     197  5.1   92     9   2
126    73     183  2.8   93     9   3
127    91     189  4.6   93     9   4
128    47      95  7.4   87     9   5
129    32      92 15.5   84     9   6
130    20     252 10.9   80     9   7
131    23     220 10.3   78     9   8
132    21     230 10.9   75     9   9
133    24     259  9.7   73     9  10
134    44     236 14.9   81     9  11
135    21     259 15.5   76     9  12
136    28     238  6.3   77     9  13
137     9      24 10.9   71     9  14
138    13     112 11.5   71     9  15
139    46     237  6.9   78     9  16
140    18     224 13.8   67     9  17
141    13      27 10.3   76     9  18
142    24     238 10.3   68     9  19
143    16     201  8.0   82     9  20
144    13     238 12.6   64     9  21
145    23      14  9.2   71     9  22
146    36     139 10.3   81     9  23
147     7      49 10.3   69     9  24
148    14      20 16.6   63     9  25
149    30     193  6.9   70     9  26
151    14     191 14.3   75     9  28
152    18     131  8.0   76     9  29
153    20     223 11.5   68     9  30
>
```

· Print the records without missing values in the dataset airquality using the function

Answer:

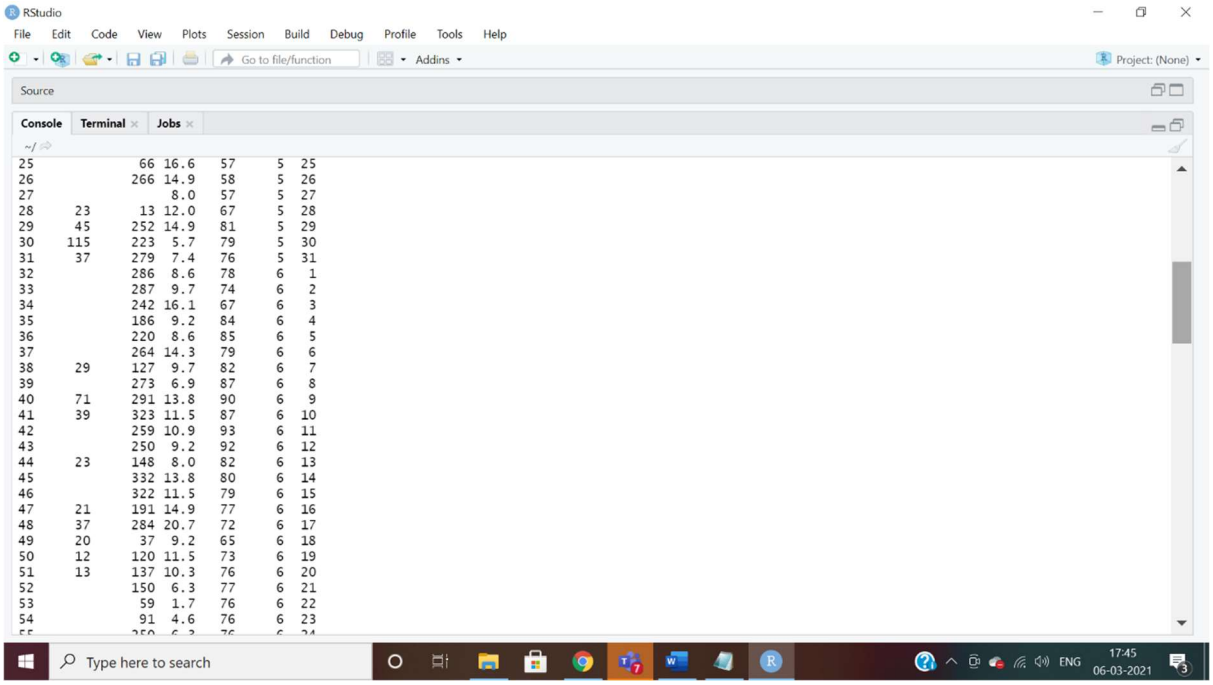
airquality[is.na(airquality)] = " "

OUTPUT



```
> #question-7
> airquality[is.na(airquality)] = " "

> airquality
   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                14.3   56     5   5
6     28          14.9   66     5   6
7     23     299  8.6   65     5   7
8     19      99 13.8   59     5   8
9      8      19 20.1   61     5   9
10            194  8.6   69     5  10
11     7           6.9   74     5  11
12    16     256  9.7   69     5  12
13    11     290  9.2   66     5  13
14    14     274 10.9   68     5  14
15    18      65 13.2   58     5  15
16    14     334 11.5   64     5  16
17    34     307 12.0   66     5  17
18     6      78 18.4   57     5  18
19    30     322 11.5   68     5  19
20    11      44  9.7   62     5  20
21     1       8  9.7   59     5  21
22    11     320 16.6   73     5  22
23     4      25  9.7   61     5  23
24    32      92 12.0   61     5  24
25            66 16.6   57     5  25
```



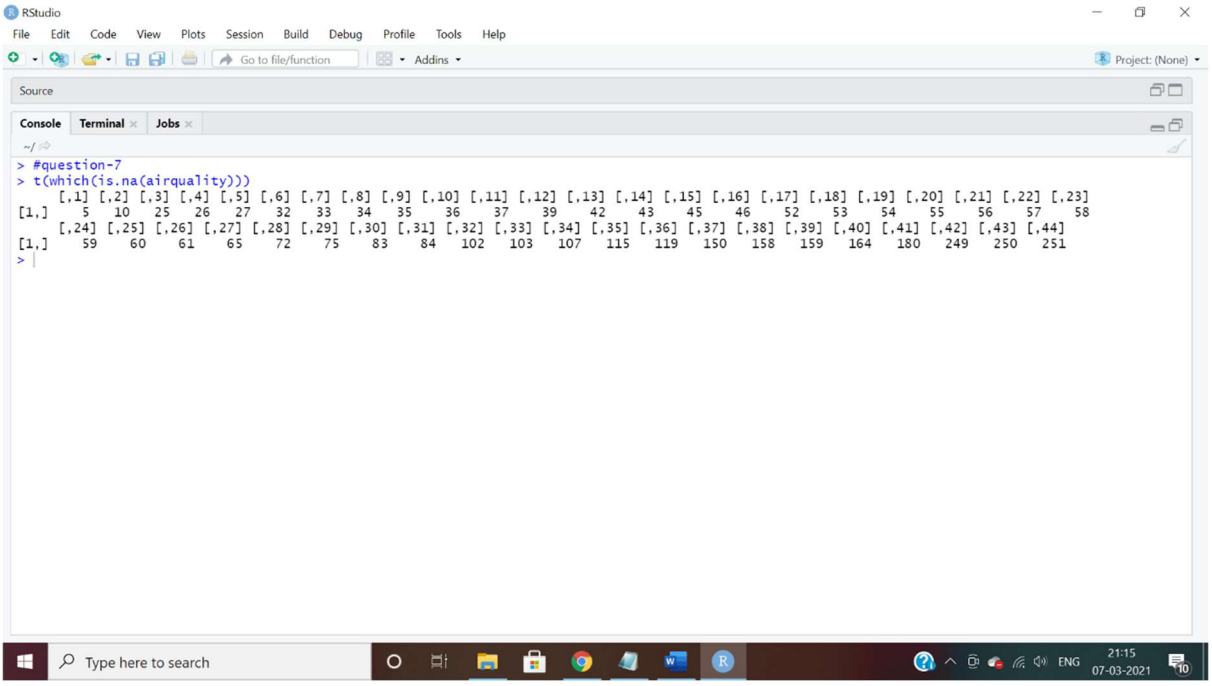
```
25            66 16.6   57     5  25
26           266 14.9   58     5  26
27                8.0   57     5  27
28    23      13 12.0   67     5  28
29    45     252 14.9   81     5  29
30   115     223  5.7   79     5  30
31    37     279  7.4   76     5  31
32           286  8.6   78     6   1
33           287  9.7   74     6   2
34           242 16.1   67     6   3
35           186  9.2   84     6   4
36           220  8.6   85     6   5
37           264 14.3   79     6   6
38    29     127  9.7   82     6   7
39           273  6.9   87     6   8
40    71     291 13.8   90     6   9
41    39     323 11.5   87     6  10
42           259 10.9   93     6  11
43           250  9.2   92     6  12
44    23     148  8.0   82     6  13
45           332 13.8   80     6  14
46           322 11.5   79     6  15
47    21     191 14.9   77     6  16
48    37     284 20.7   72     6  17
49    20      37  9.2   65     6  18
50    12     120 11.5   73     6  19
51    13     137 10.3   76     6  20
52           150  6.3   77     6  21
53            59  1.7   76     6  22
54            91  4.6   76     6  23
```

· Print the indices of the missing values in the dataset airquality in column major representation

Answer:

t(which(is.na(airquality)))

OUTPUT



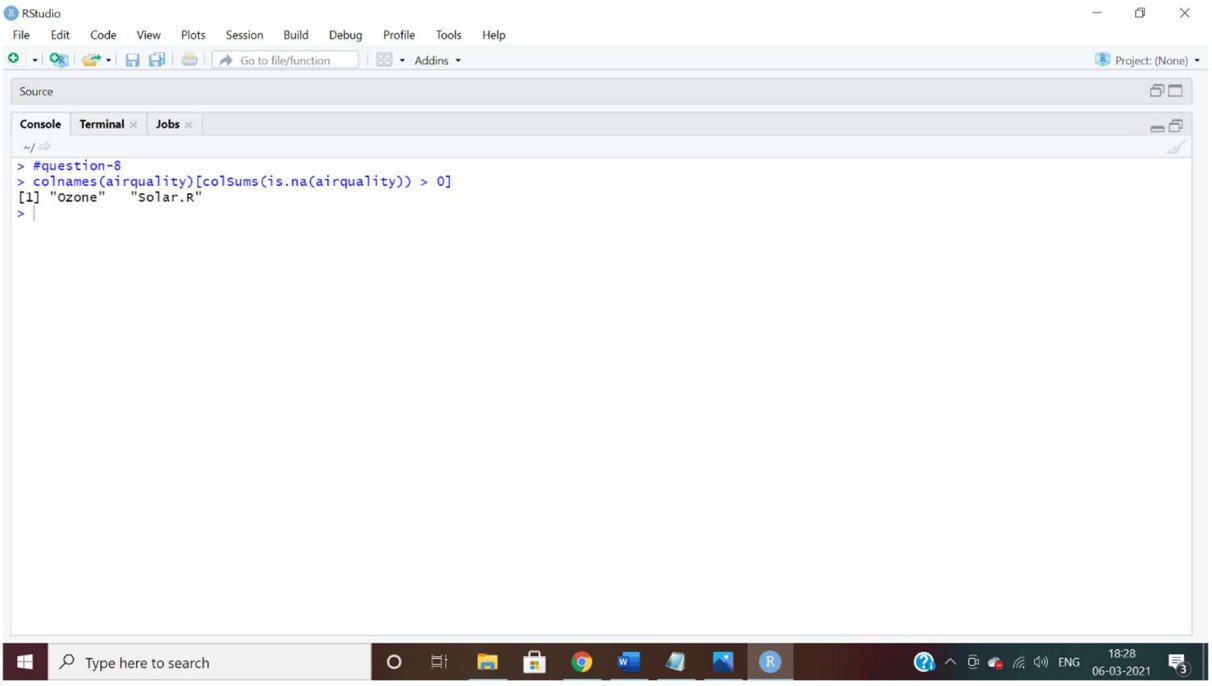
```
> #question-7
> t(which(is.na(airquality)))
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23]
[1,]    5   10   25   26   27   32   33   34   35    36    37    39    42    43    45    46    52    53    54    55    56    57    58
     [,24] [,25] [,26] [,27] [,28] [,29] [,30] [,31] [,32] [,33] [,34] [,35] [,36] [,37] [,38] [,39] [,40] [,41] [,42] [,43] [,44]
[1,]    59    60    61    65    72    75    83    84   102   103   107   115   119   150   158   159   164   180   249   250   251
>
```

· **Print Names of the Columns which contains Missing Values in the dataset airquality**

Answer:

colnames(airquality)[colSums(is.na(airquality)) > 0]

OUTPUT



Data Recording

Consider a numeric vector x <- c(3,4,5,6,7,8)

Write a command to recode the values less than 6 with zero in the vector x

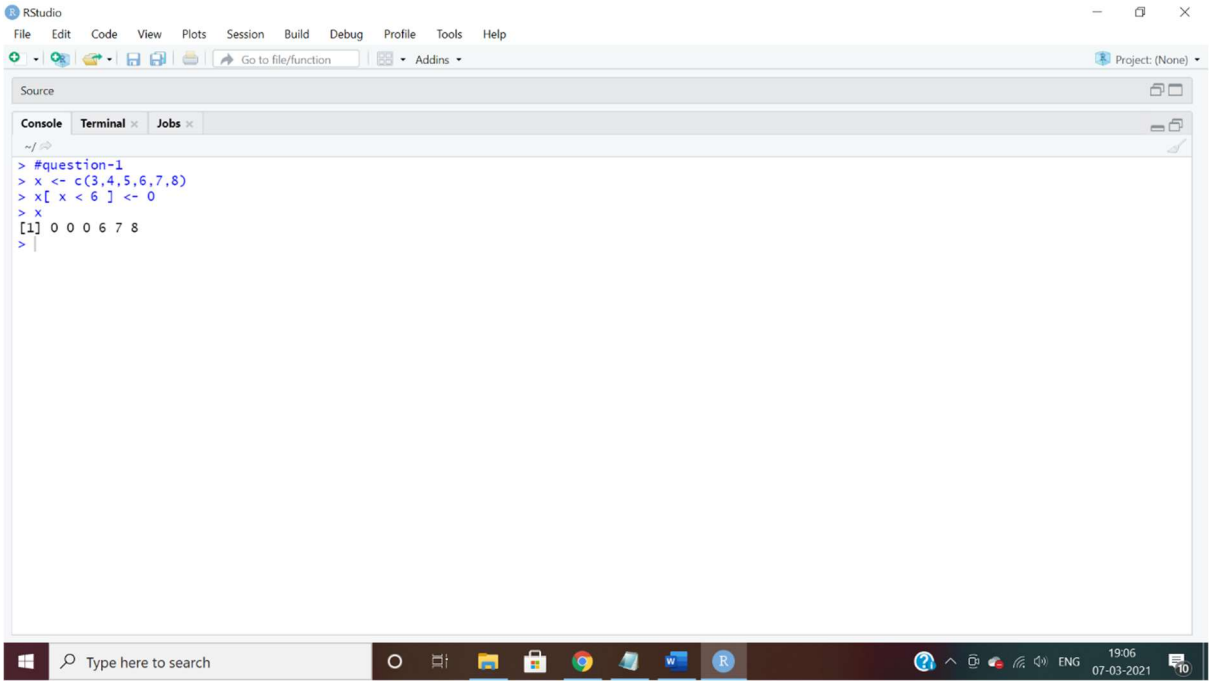
Answer:

x <- c(3,4,5,6,7,8)

x[x < 6] <- 0

x

OUTPUT



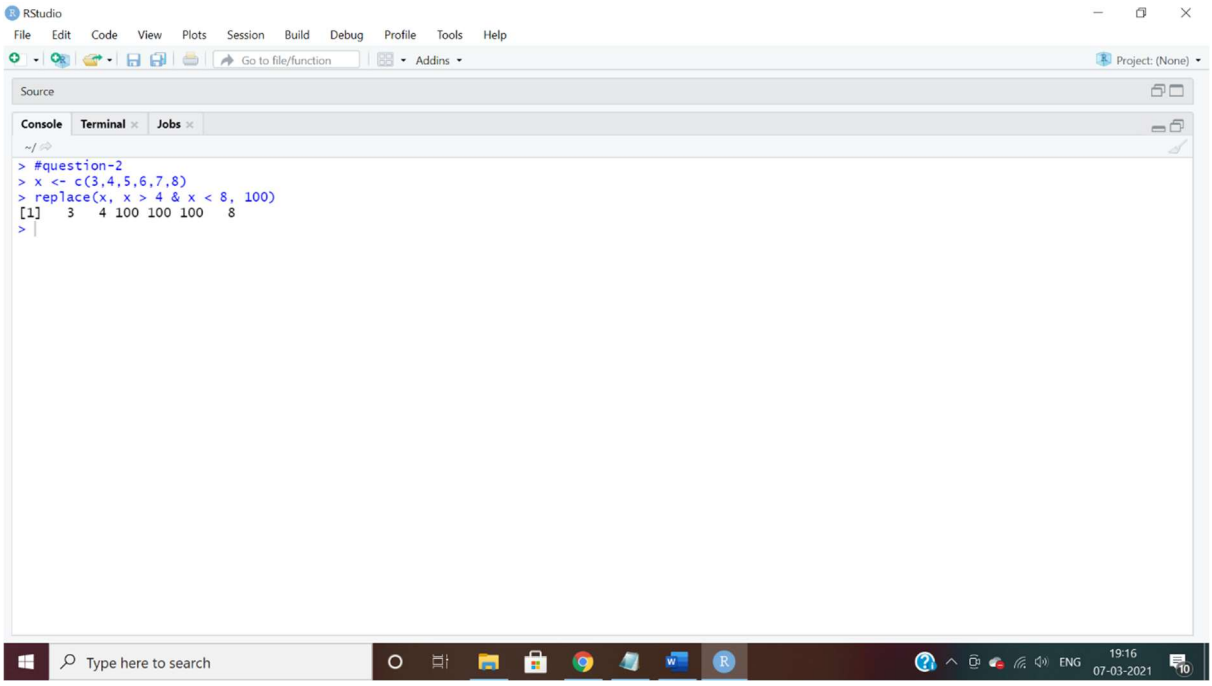
- **Write a command to recode the values between 4 and 8 with 100**

Answer:

x <- c(3,4,5,6,7,8)

replace(x, x > 4 & x < 8, 100)

OUTPUT



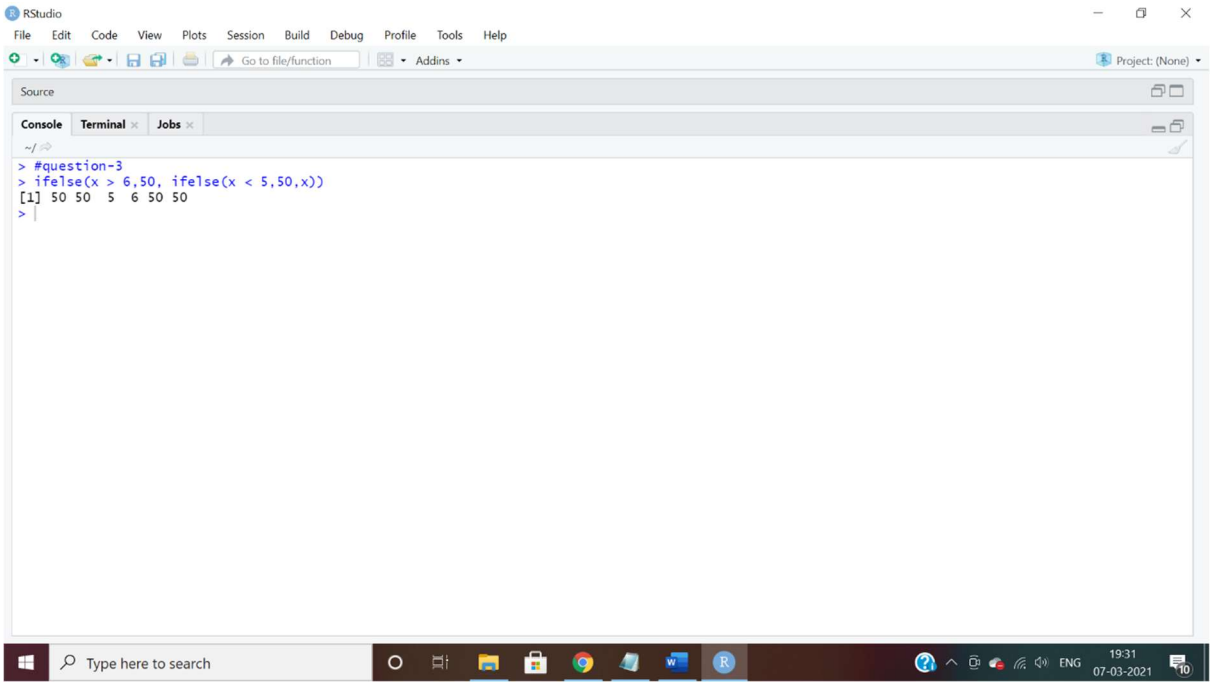
- **Write a command to recode the values that are less than 5 or greater than 6 with 50**

Answer:

x <- c(3,4,5,6,7,8)

ifelse(x > 6,50, ifelse(x < 5,50,x))

OUTPUT



- **Write a command to recode the values less than 6 with NA in the vector x**

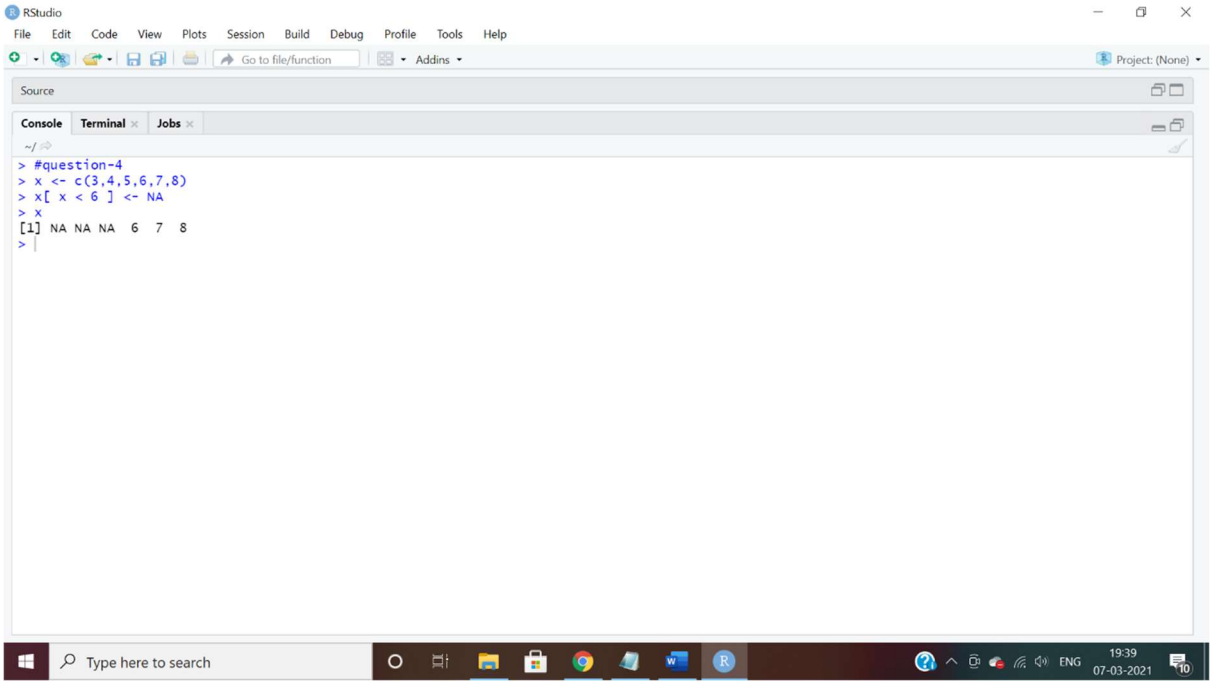
Answer:

x <- c(3,4,5,6,7,8)

x[x < 6] <- NA

x

OUTPUT



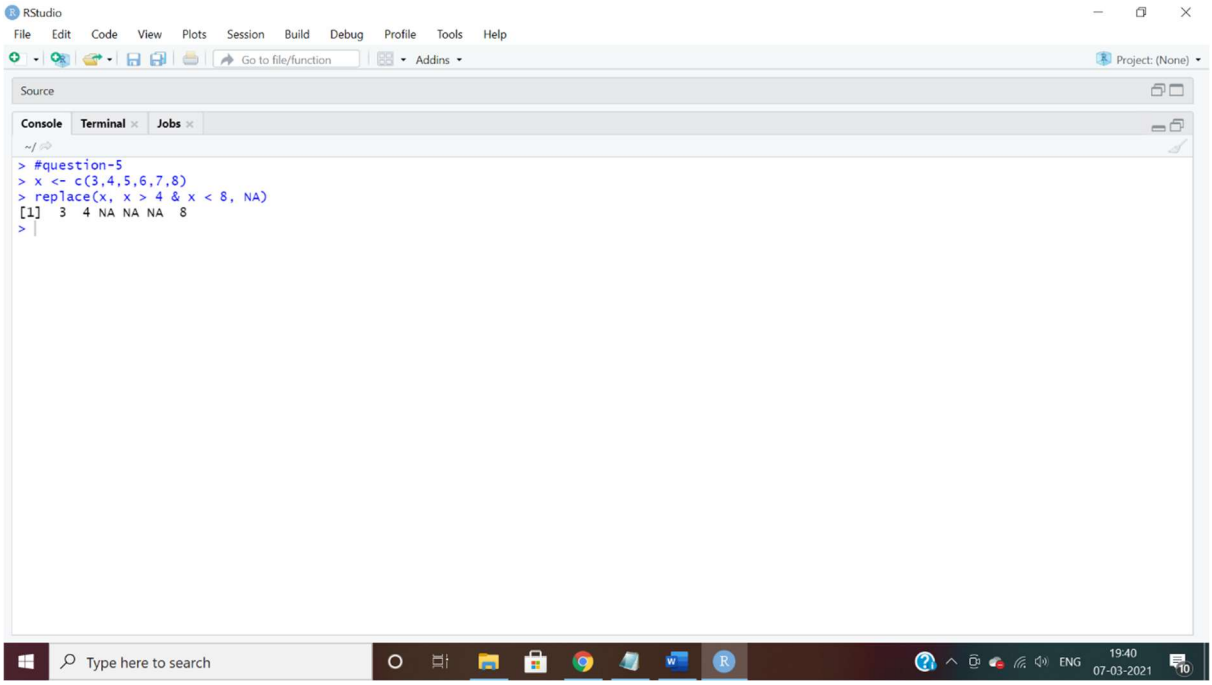
- **Write a command to recode the values between 4 and 8 with NA**

Answer:

x <- c(3,4,5,6,7,8)

replace(x, x > 4 & x < 8, NA)

OUTPUT



- **Write a command to recode the values that are less than 5 or greater than 6 with NA**

Answer:

x <- c(3,4,5,6,7,8)

ifelse(x > 6,NA, ifelse(x < 5,NA,x))

OUTPUT