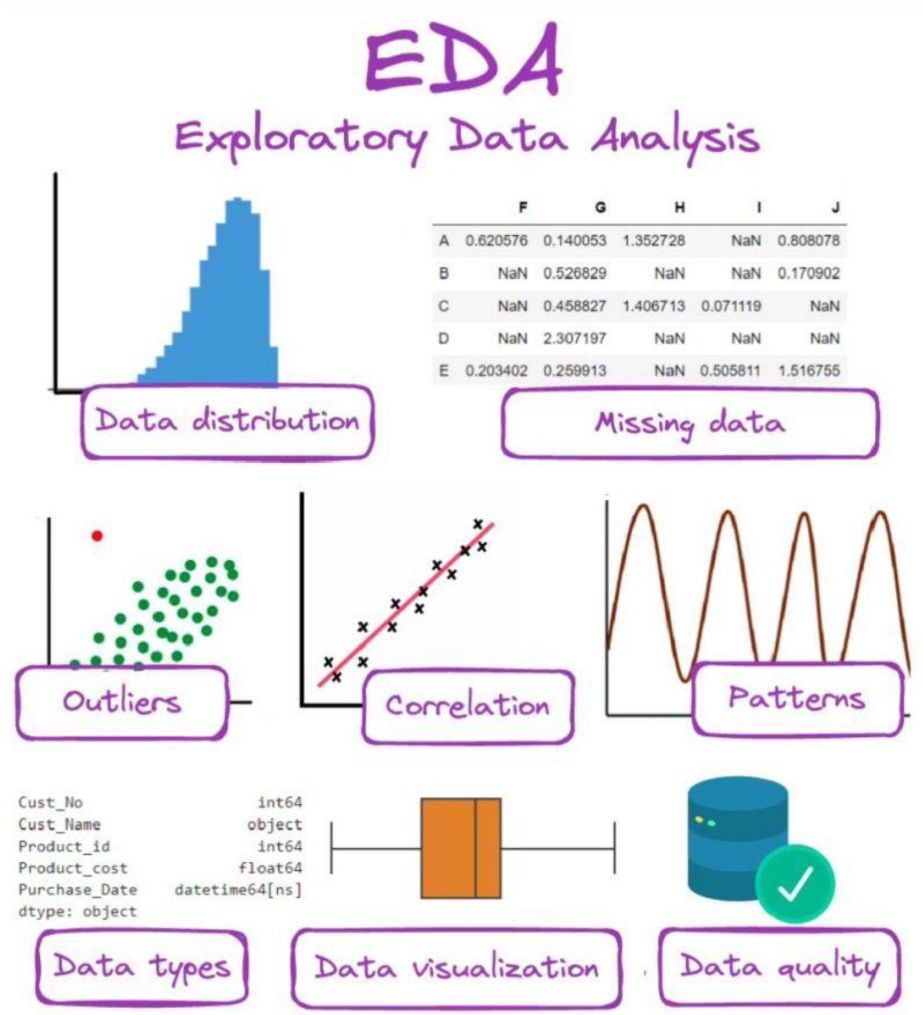
EXPERIMENT – 5

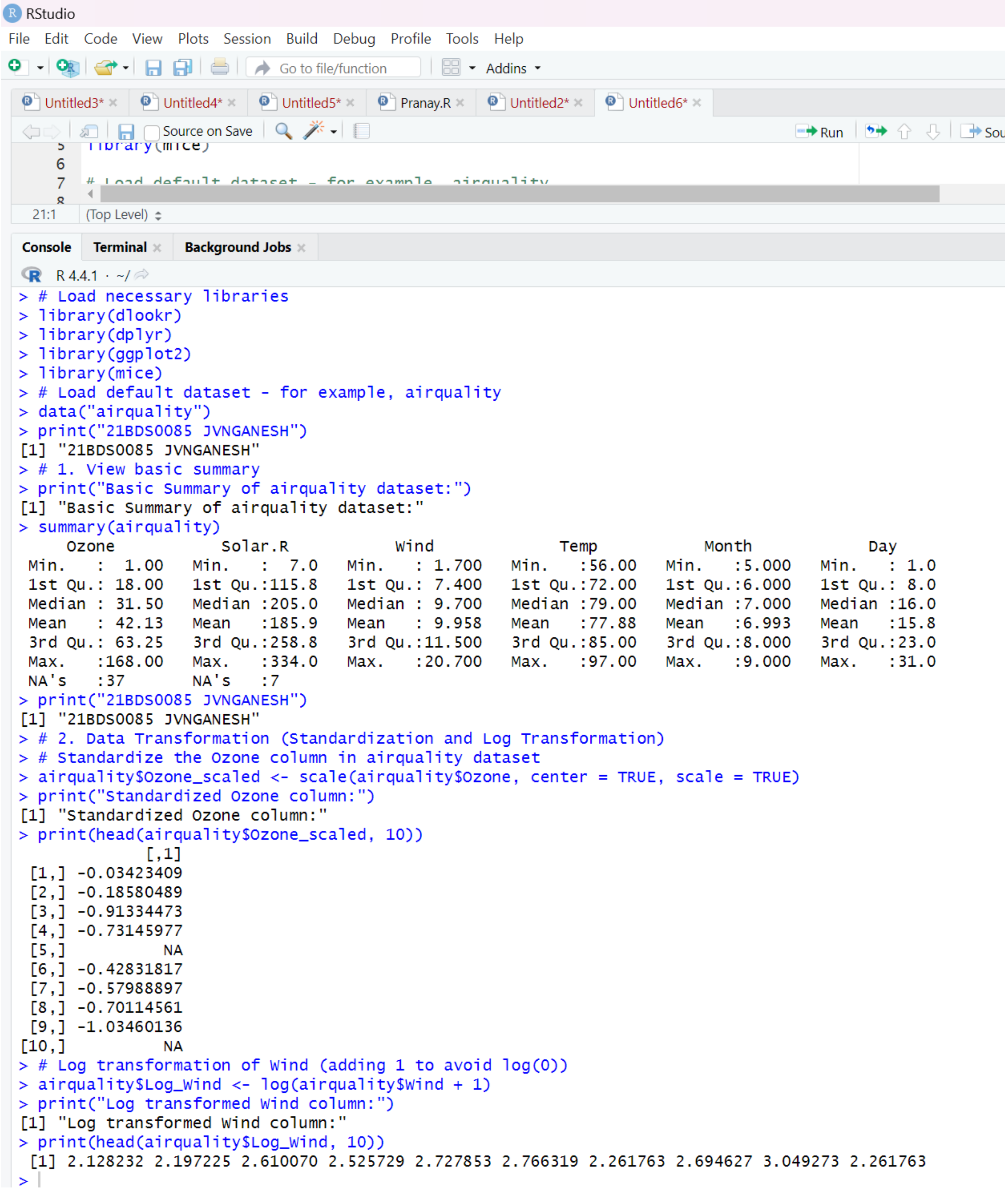


	F	G	H	I	J
A	0.620576	0.140053	1.352728	NaN	0.808078
B	NaN	0.526829	NaN	NaN	0.170902
C	NaN	0.458827	1.406713	0.071119	NaN
D	NaN	2.307197	NaN	NaN	NaN
E	0.203402	0.259913	NaN	0.505811	1.516755

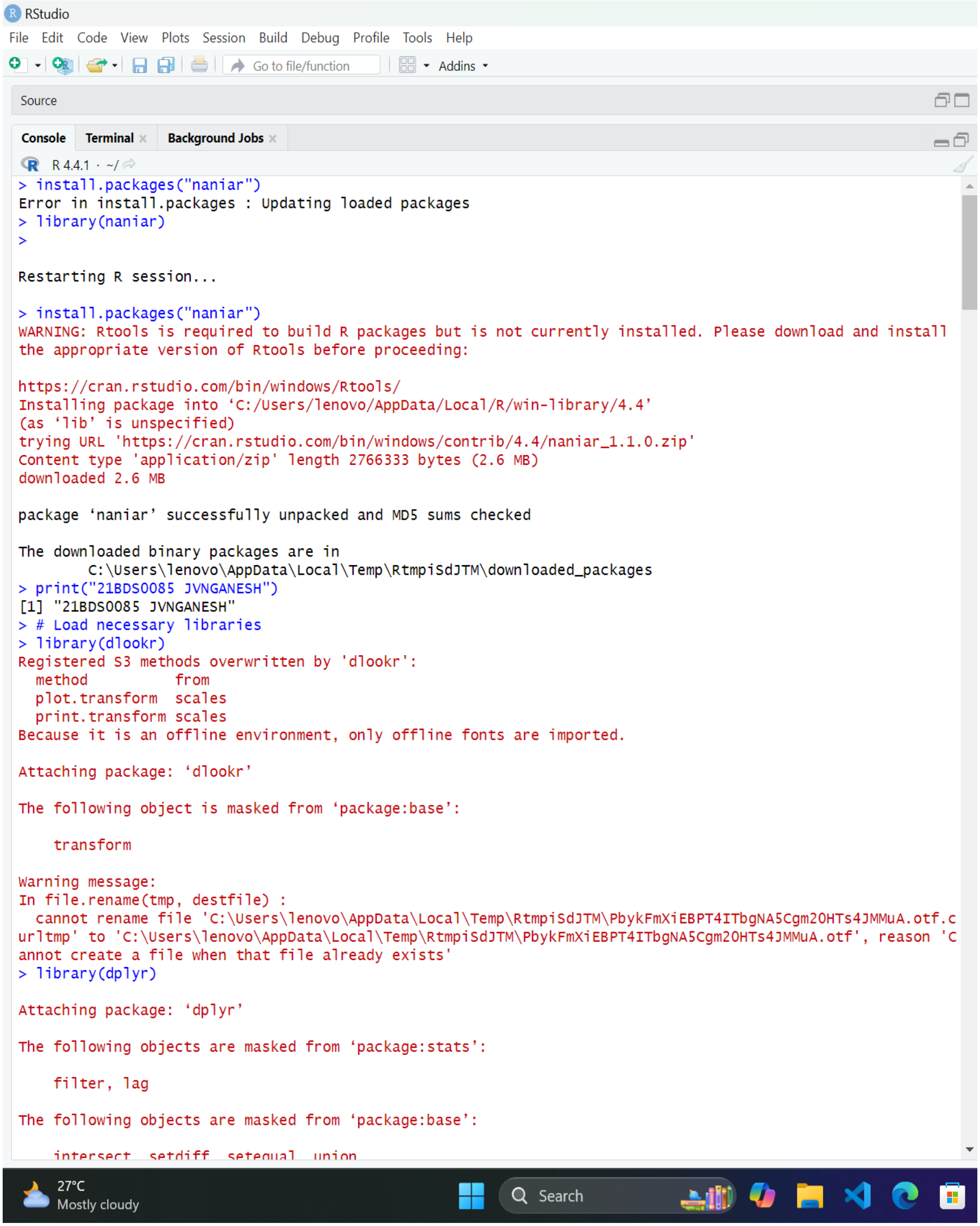
```
Cust_No                  int64
Cust_Name               object
Product_id               int64
Product_cost           float64
Purchase_Date   datetime64[ns]
dtype: object
```

NAME : JVN GANESH

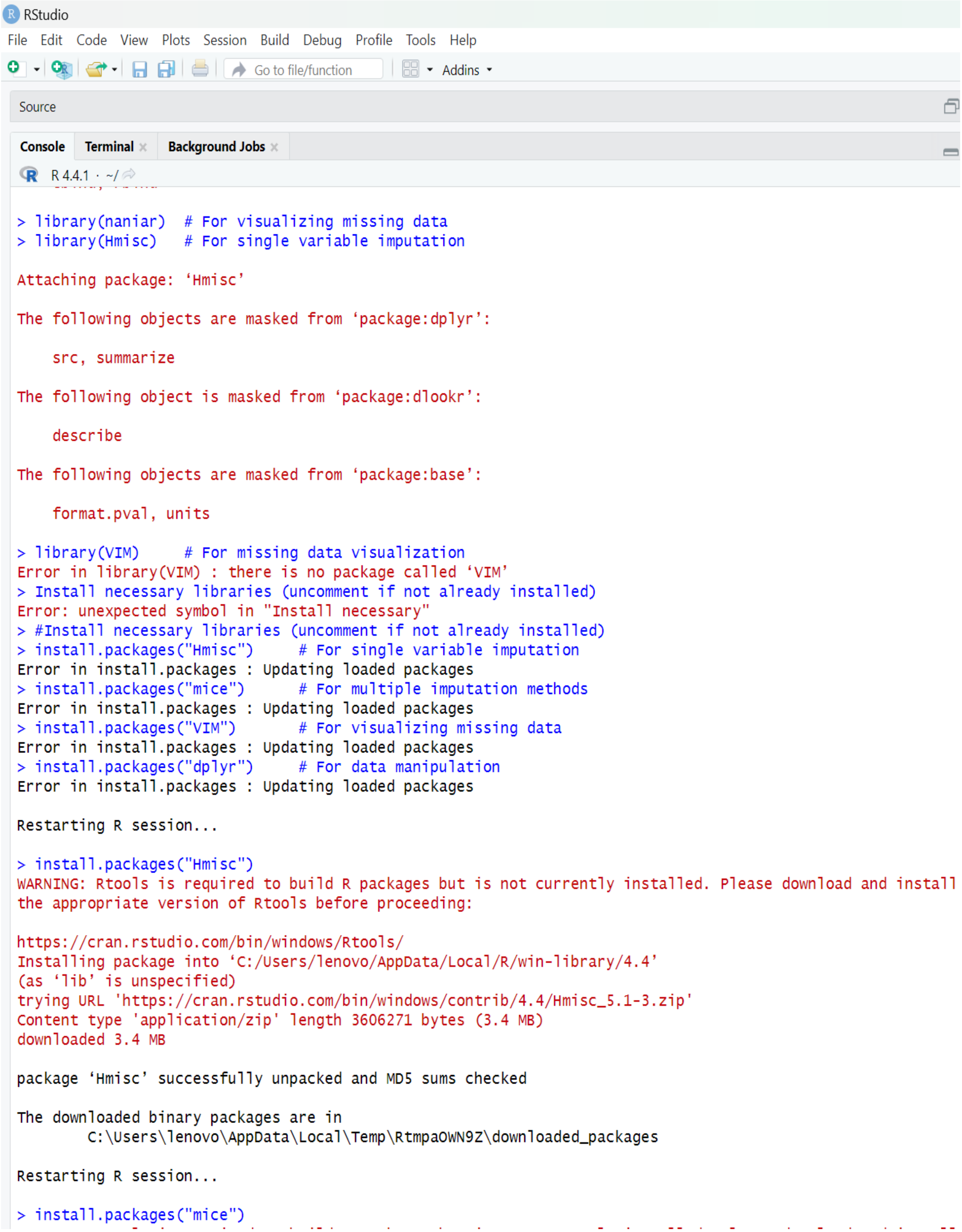
Roll No : 21BDS0085



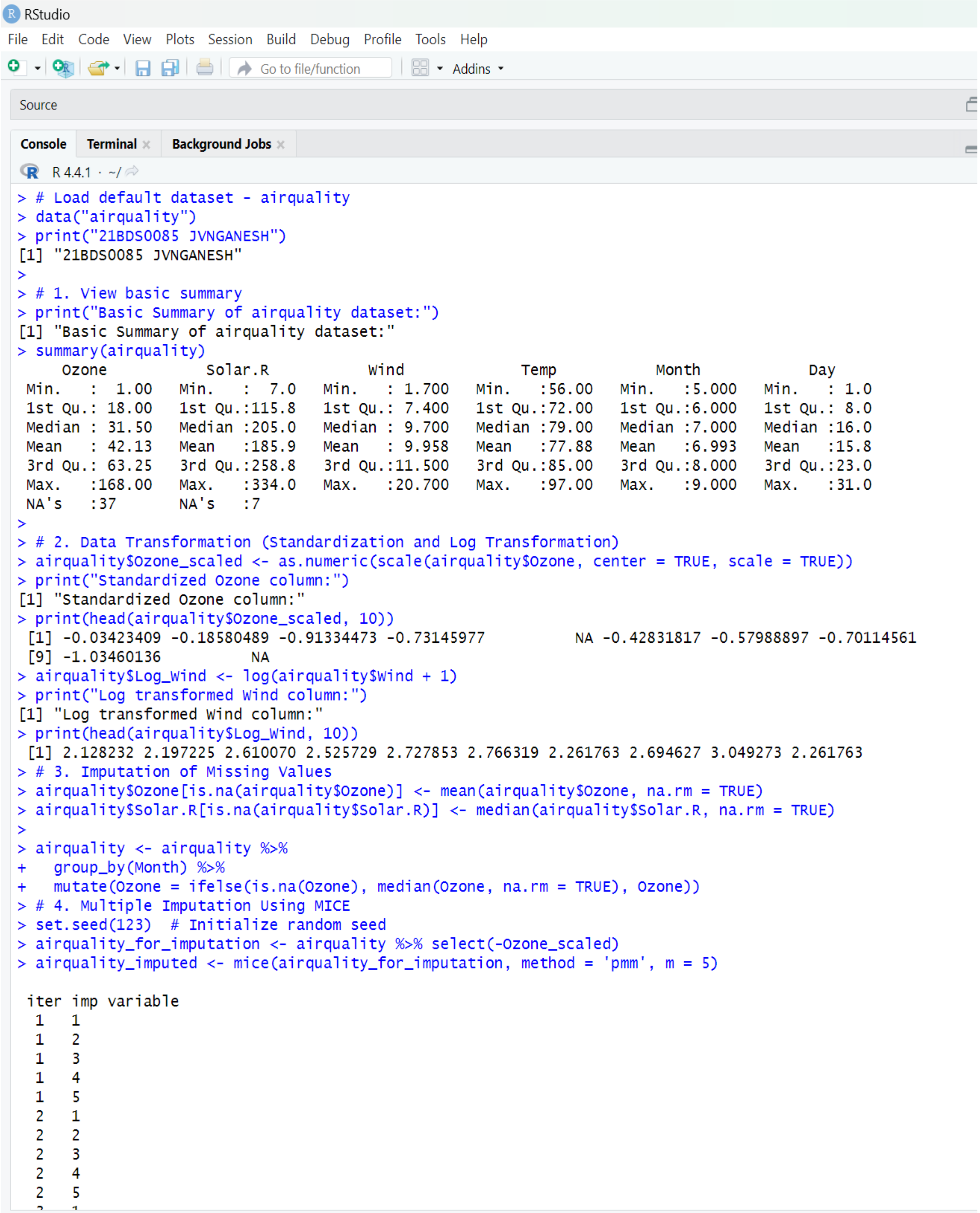
```
> # Load necessary libraries
> library(dlookr)
> library(dplyr)
> library(ggplot2)
> library(mice)
> # Load default dataset - for example, airquality
> data("airquality")
> print("21BDS0085 JVNGANESH")
[1] "21BDS0085 JVNGANESH"
> # 1. View basic summary
> print("Basic Summary of airquality dataset:")
[1] "Basic Summary of airquality dataset:"
> summary(airquality)
     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                                                                       
> print("21BDS0085 JVNGANESH")
[1] "21BDS0085 JVNGANESH"
> # 2. Data Transformation (Standardization and Log Transformation)
> # Standardize the Ozone column in airquality dataset
> airquality$Ozone_scaled <- scale(airquality$Ozone, center = TRUE, scale = TRUE)
> print("Standardized Ozone column:")
[1] "Standardized Ozone column:"
> print(head(airquality$Ozone_scaled, 10))
             [,1]
 [1,] -0.03423409
 [2,] -0.18580489
 [3,] -0.91334473
 [4,] -0.73145977
 [5,]          NA
 [6,] -0.42831817
 [7,] -0.57988897
 [8,] -0.70114561
 [9,] -1.03460136
[10,]          NA
> # Log transformation of Wind (adding 1 to avoid log(0))
> airquality$Log_Wind <- log(airquality$Wind + 1)
> print("Log transformed Wind column:")
[1] "Log transformed Wind column:"
> print(head(airquality$Log_Wind, 10))
 [1] 2.128232 2.197225 2.610070 2.525729 2.727853 2.766319 2.261763 2.694627 3.049273 2.261763
> 
```



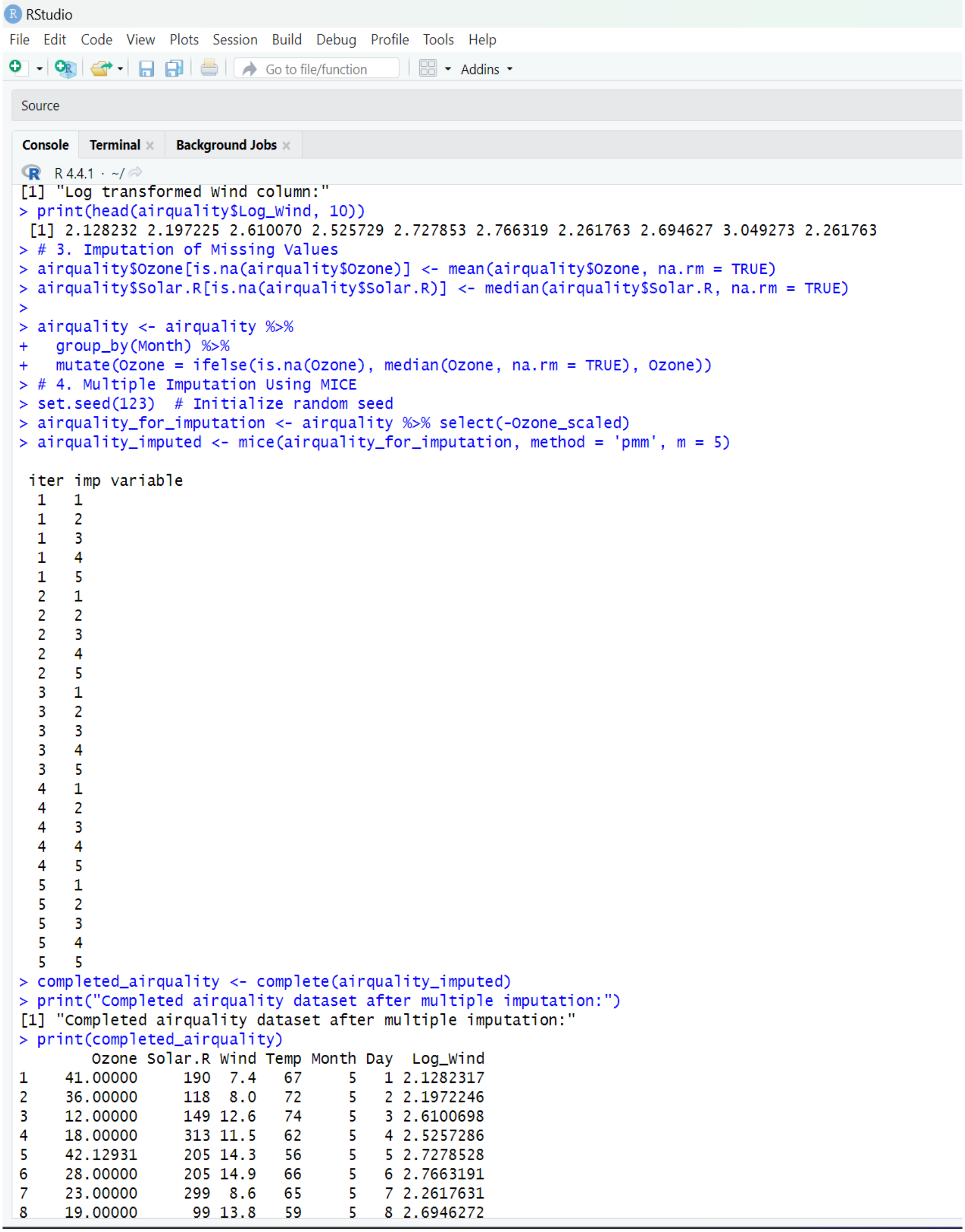
```
> install.packages("naniar")
Error in install.packages : Updating loaded packages
> library(naniar)
>

Restarting R session...

> install.packages("naniar")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install
the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/lenovo/AppData/Local/R/win-library/4.4’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.4/naniar_1.1.0.zip'
Content type 'application/zip' length 2766333 bytes (2.6 MB)
downloaded 2.6 MB

package ‘naniar’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\lenovo\AppData\Local\Temp\RtmpiSdJTM\downloaded_packages
> print("21BDS0085 JVNGANESH")
[1] "21BDS0085 JVNGANESH"
> # Load necessary libraries
> library(dlookr)
Registered S3 methods overwritten by 'dlookr':
  method          from
  plot.transform  scales
  print.transform scales
Because it is an offline environment, only offline fonts are imported.

Attaching package: ‘dlookr’

The following object is masked from ‘package:base’:

    transform

Warning message:
In file.rename(tmp, destfile) :
  cannot rename file 'C:\Users\lenovo\AppData\Local\Temp\RtmpiSdJTM\PbykFmXiEBPT4ITbgNA5Cgm20HTs4JMMuA.otf.c
urltmp' to 'C:\Users\lenovo\AppData\Local\Temp\RtmpiSdJTM\PbykFmXiEBPT4ITbgNA5Cgm20HTs4JMMuA.otf', reason 'C
annot create a file when that file already exists'
> library(dplyr)

Attaching package: ‘dplyr’

The following objects are masked from ‘package:stats’:

    filter, lag

The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union
```



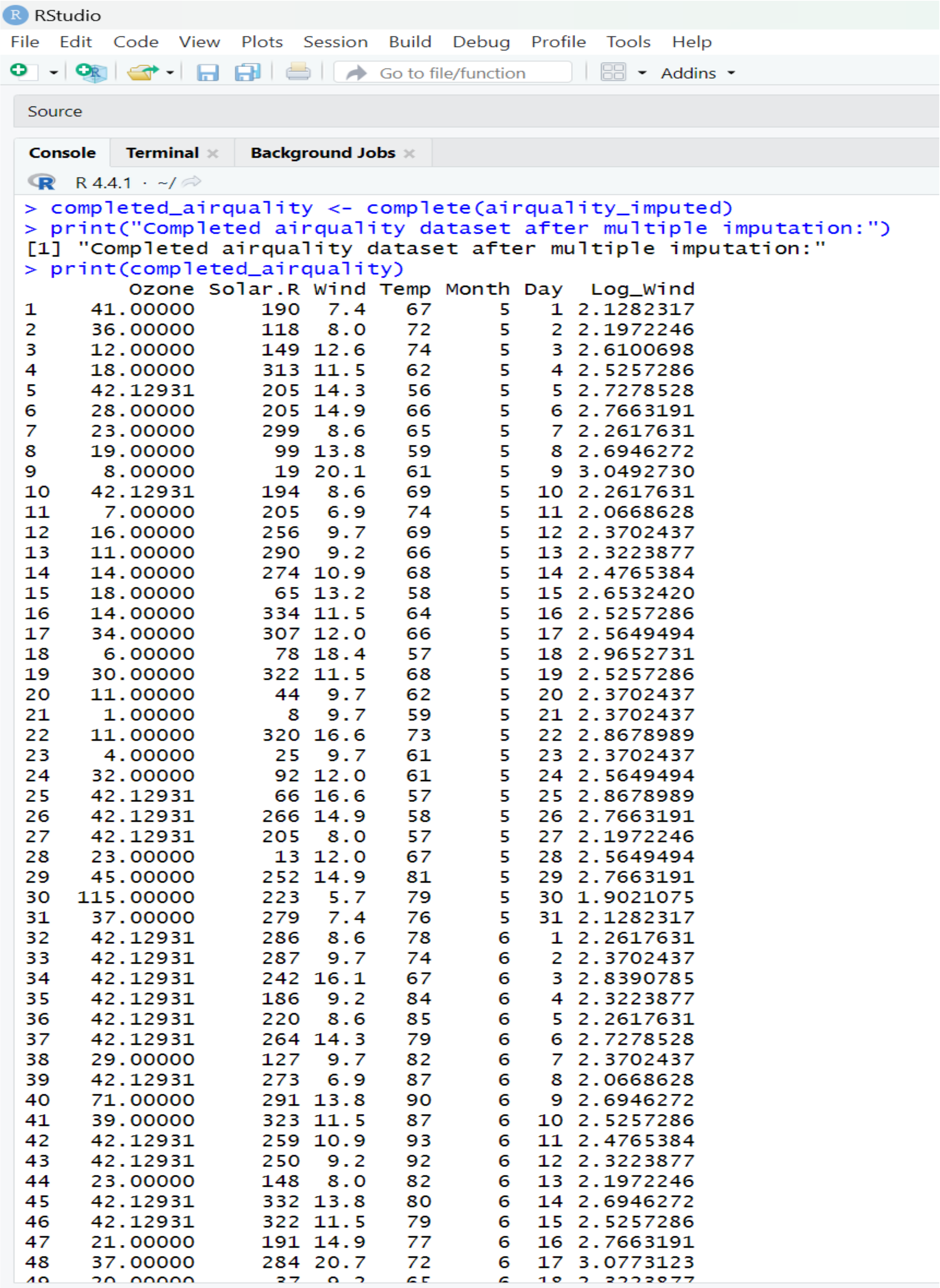
```
> library(naniar)  # For visualizing missing data
> library(Hmisc)   # For single variable imputation

Attaching package: ‘Hmisc’

The following objects are masked from ‘package:dplyr’:

    src, summarize

The following object is masked from ‘package:dlookr’:

    describe

The following objects are masked from ‘package:base’:

    format.pval, units

> library(VIM)     # For missing data visualization
Error in library(VIM) : there is no package called ‘VIM’
> Install necessary libraries (uncomment if not already installed)
Error: unexpected symbol in "Install necessary"
> #Install necessary libraries (uncomment if not already installed)
> install.packages("Hmisc")     # For single variable imputation
Error in install.packages : Updating loaded packages
> install.packages("mice")      # For multiple imputation methods
Error in install.packages : Updating loaded packages
> install.packages("VIM")       # For visualizing missing data
Error in install.packages : Updating loaded packages
> install.packages("dplyr")     # For data manipulation
Error in install.packages : Updating loaded packages

Restarting R session...

> install.packages("Hmisc")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install
the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into ‘C:/Users/lenovo/AppData/Local/R/win-library/4.4’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.4/Hmisc_5.1-3.zip'
Content type 'application/zip' length 3606271 bytes (3.4 MB)
downloaded 3.4 MB

package ‘Hmisc’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\lenovo\AppData\Local\Temp\RtmpaOWN9Z\downloaded_packages

Restarting R session...

> install.packages("mice")
```



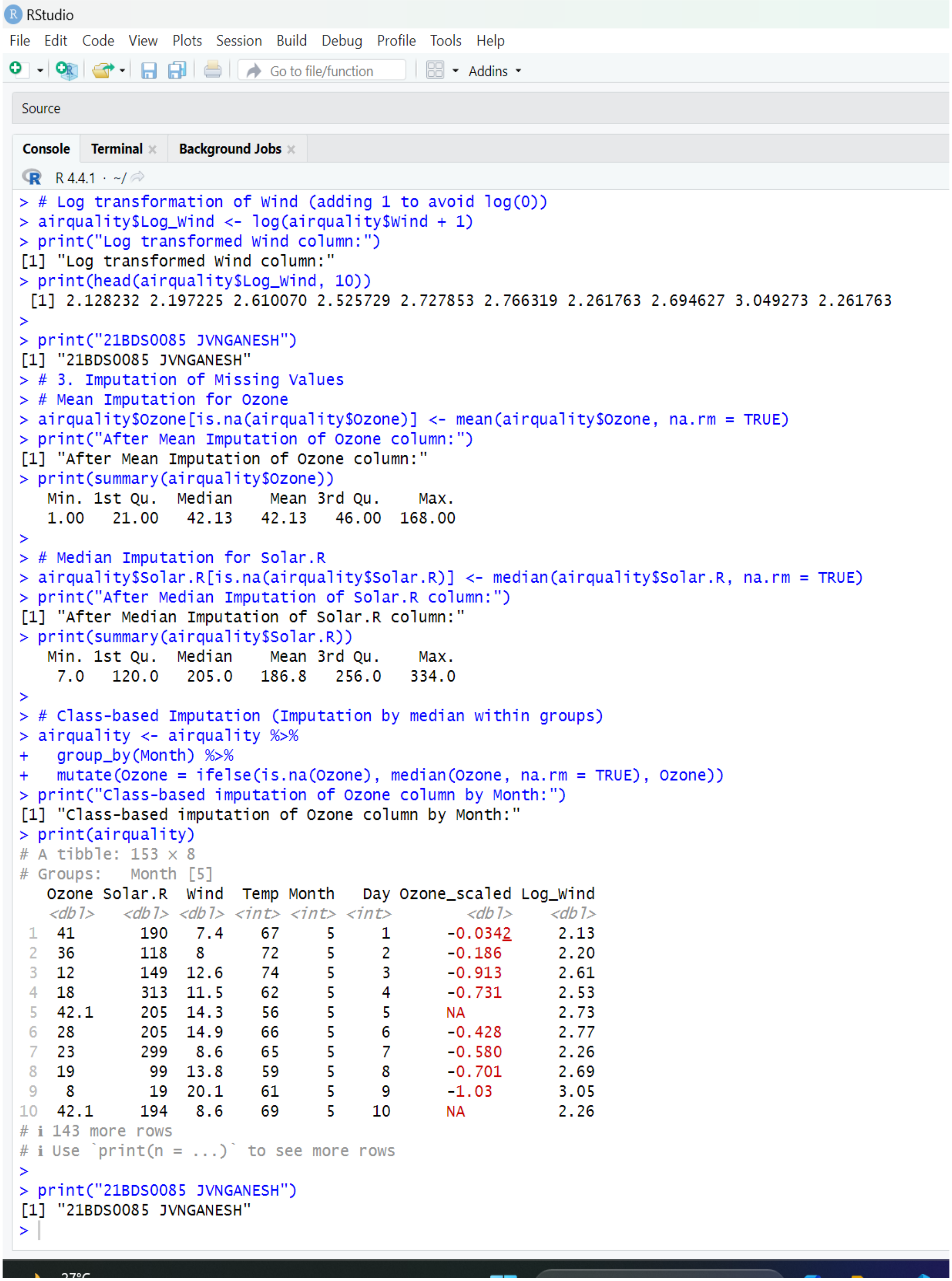
```
> # Load default dataset - airquality
> data("airquality")
> print("21BDS0085 JVNGANESH")
[1] "21BDS0085 JVNGANESH"
> 
> # 1. View basic summary
> print("Basic Summary of airquality dataset:")
[1] "Basic Summary of airquality dataset:"
> summary(airquality)
     Ozone           Solar.R           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                                                                       
> 
> # 2. Data Transformation (Standardization and Log Transformation)
> airquality$Ozone_scaled <- as.numeric(scale(airquality$Ozone, center = TRUE, scale = TRUE))
> print("Standardized Ozone column:")
[1] "Standardized Ozone column:"
> print(head(airquality$Ozone_scaled, 10))
 [1] -0.03423409 -0.18580489 -0.91334473 -0.73145977          NA -0.42831817 -0.57988897 -0.70114561
 [9] -1.03460136          NA
> airquality$Log_Wind <- log(airquality$Wind + 1)
> print("Log transformed Wind column:")
[1] "Log transformed Wind column:"
> print(head(airquality$Log_Wind, 10))
 [1] 2.128232 2.197225 2.610070 2.525729 2.727853 2.766319 2.261763 2.694627 3.049273 2.261763
> # 3. Imputation of Missing Values
> airquality$Ozone[is.na(airquality$Ozone)] <- mean(airquality$Ozone, na.rm = TRUE)
> airquality$Solar.R[is.na(airquality$Solar.R)] <- median(airquality$Solar.R, na.rm = TRUE)
> 
> airquality <- airquality %>%
+   group_by(Month) %>%
+   mutate(Ozone = ifelse(is.na(Ozone), median(Ozone, na.rm = TRUE), Ozone))
> # 4. Multiple Imputation Using MICE
> set.seed(123)  # Initialize random seed
> airquality_for_imputation <- airquality %>% select(-Ozone_scaled)
> airquality_imputed <- mice(airquality_for_imputation, method = 'pmm', m = 5)

 iter imp variable
  1   1
  1   2
  1   3
  1   4
  1   5
  2   1
  2   2
  2   3
  2   4
  2   5
```



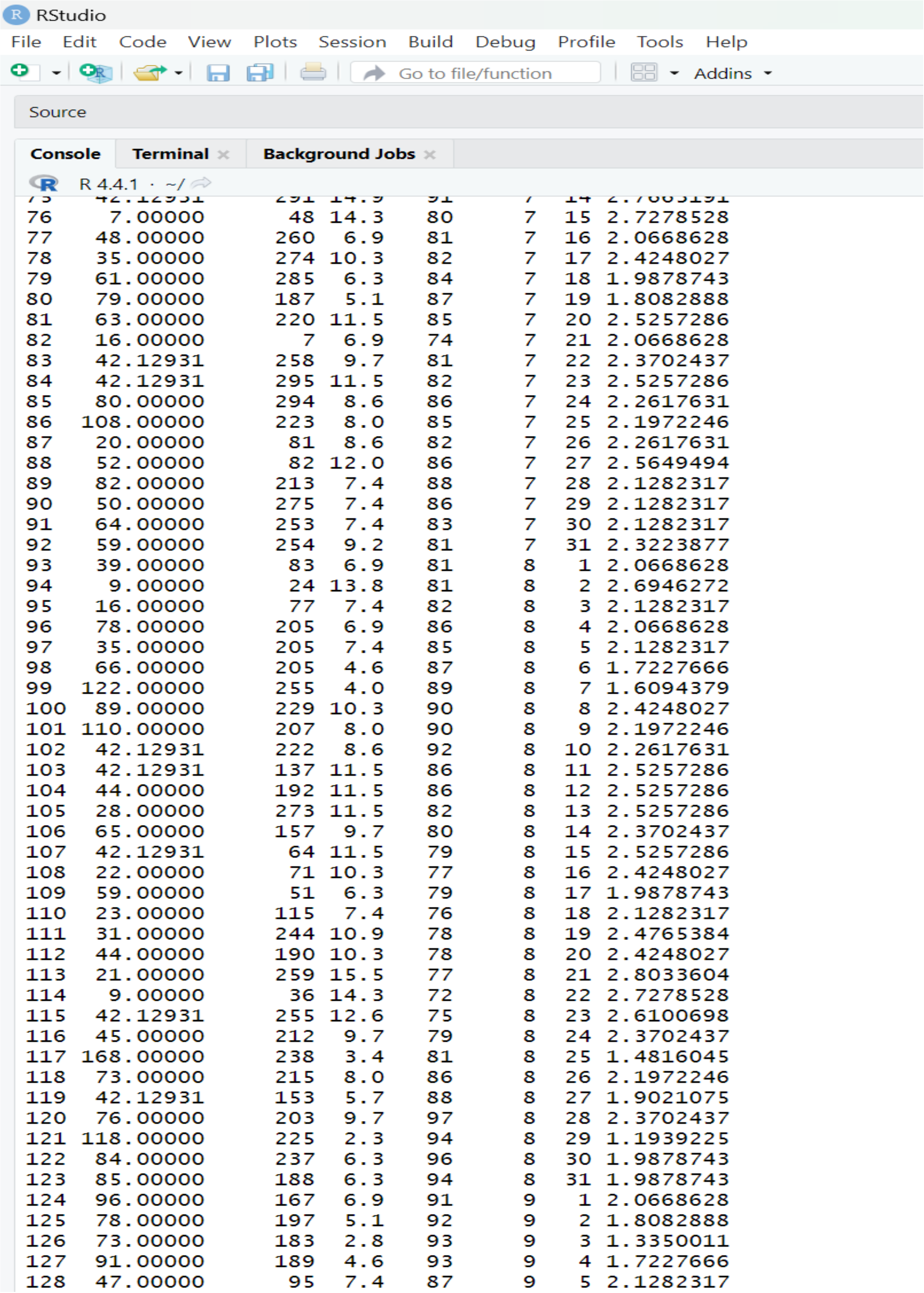
```
[1] "Log transformed Wind column:"
> print(head(airquality$Log_Wind, 10))
 [1] 2.128232 2.197225 2.610070 2.525729 2.727853 2.766319 2.261763 2.694627 3.049273 2.261763
> # 3. Imputation of Missing Values
> airquality$Ozone[is.na(airquality$Ozone)] <- mean(airquality$Ozone, na.rm = TRUE)
> airquality$Solar.R[is.na(airquality$Solar.R)] <- median(airquality$Solar.R, na.rm = TRUE)
> 
> airquality <- airquality %>%
+   group_by(Month) %>%
+   mutate(Ozone = ifelse(is.na(Ozone), median(Ozone, na.rm = TRUE), Ozone))
> # 4. Multiple Imputation Using MICE
> set.seed(123)  # Initialize random seed
> airquality_for_imputation <- airquality %>% select(-Ozone_scaled)
> airquality_imputed <- mice(airquality_for_imputation, method = 'pmm', m = 5)

 iter imp variable
  1   1
  1   2
  1   3
  1   4
  1   5
  2   1
  2   2
  2   3
  2   4
  2   5
  3   1
  3   2
  3   3
  3   4
  3   5
  4   1
  4   2
  4   3
  4   4
  4   5
  5   1
  5   2
  5   3
  5   4
  5   5
> completed_airquality <- complete(airquality_imputed)
> print("Completed airquality dataset after multiple imputation:")
[1] "Completed airquality dataset after multiple imputation:"
> print(completed_airquality)
       Ozone Solar.R Wind Temp Month Day  Log_Wind
1   41.00000     190  7.4   67     5   1 2.1282317
2   36.00000     118  8.0   72     5   2 2.1972246
3   12.00000     149 12.6   74     5   3 2.6100698
4   18.00000     313 11.5   62     5   4 2.5257286
5   42.12931     205 14.3   56     5   5 2.7278528
6   28.00000     205 14.9   66     5   6 2.7663191
7   23.00000     299  8.6   65     5   7 2.2617631
8   19.00000      99 13.8   59     5   8 2.6946272
```



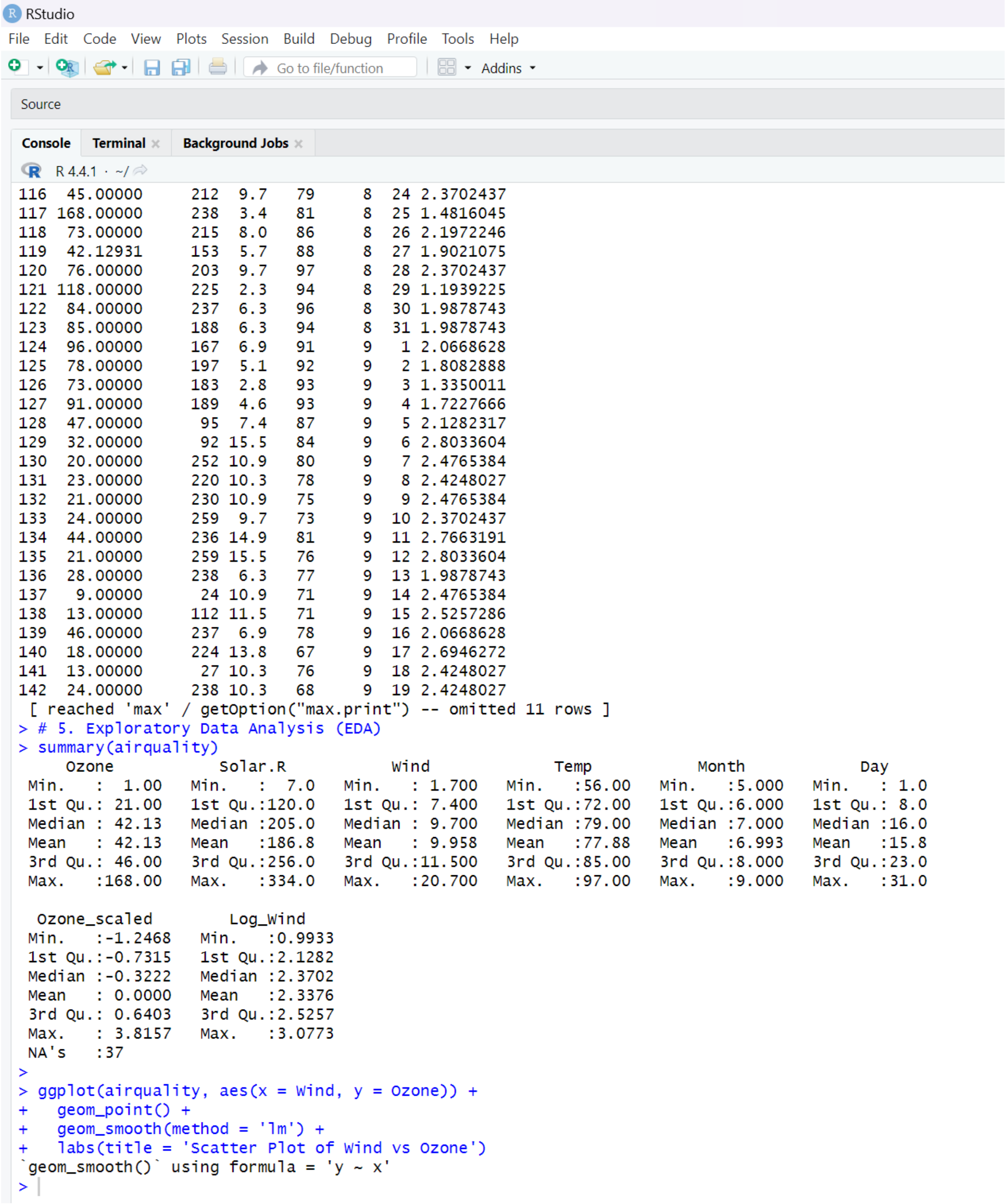
```
> completed_airquality <- complete(airquality_imputed)
> print("Completed airquality dataset after multiple imputation:")
[1] "Completed airquality dataset after multiple imputation:"
> print(completed_airquality)
        Ozone Solar.R Wind Temp Month Day  Log_Wind
1    41.00000     190  7.4   67     5   1 2.1282317
2    36.00000     118  8.0   72     5   2 2.1972246
3    12.00000     149 12.6   74     5   3 2.6100698
4    18.00000     313 11.5   62     5   4 2.5257286
5    42.12931     205 14.3   56     5   5 2.7278528
6    28.00000     205 14.9   66     5   6 2.7663191
7    23.00000     299  8.6   65     5   7 2.2617631
8    19.00000      99 13.8   59     5   8 2.6946272
9     8.00000      19 20.1   61     5   9 3.0492730
10   42.12931     194  8.6   69     5  10 2.2617631
11    7.00000     205  6.9   74     5  11 2.0668628
12   16.00000     256  9.7   69     5  12 2.3702437
13   11.00000     290  9.2   66     5  13 2.3223877
14   14.00000     274 10.9   68     5  14 2.4765384
15   18.00000      65 13.2   58     5  15 2.6532420
16   14.00000     334 11.5   64     5  16 2.5257286
17   34.00000     307 12.0   66     5  17 2.5649494
18    6.00000      78 18.4   57     5  18 2.9652731
19   30.00000     322 11.5   68     5  19 2.5257286
20   11.00000      44  9.7   62     5  20 2.3702437
21    1.00000       8  9.7   59     5  21 2.3702437
22   11.00000     320 16.6   73     5  22 2.8678989
23    4.00000      25  9.7   61     5  23 2.3702437
24   32.00000      92 12.0   61     5  24 2.5649494
25   42.12931      66 16.6   57     5  25 2.8678989
26   42.12931     266 14.9   58     5  26 2.7663191
27   42.12931     205  8.0   57     5  27 2.1972246
28   23.00000      13 12.0   67     5  28 2.5649494
29   45.00000     252 14.9   81     5  29 2.7663191
30  115.00000     223  5.7   79     5  30 1.9021075
31   37.00000     279  7.4   76     5  31 2.1282317
32   42.12931     286  8.6   78     6   1 2.2617631
33   42.12931     287  9.7   74     6   2 2.3702437
34   42.12931     242 16.1   67     6   3 2.8390785
35   42.12931     186  9.2   84     6   4 2.3223877
36   42.12931     220  8.6   85     6   5 2.2617631
37   42.12931     264 14.3   79     6   6 2.7278528
38   29.00000     127  9.7   82     6   7 2.3702437
39   42.12931     273  6.9   87     6   8 2.0668628
40   71.00000     291 13.8   90     6   9 2.6946272
41   39.00000     323 11.5   87     6  10 2.5257286
42   42.12931     259 10.9   93     6  11 2.4765384
43   42.12931     250  9.2   92     6  12 2.3223877
44   23.00000     148  8.0   82     6  13 2.1972246
45   42.12931     332 13.8   80     6  14 2.6946272
46   42.12931     322 11.5   79     6  15 2.5257286
47   21.00000     191 14.9   77     6  16 2.7663191
48   37.00000     284 20.7   72     6  17 3.0773123
49   20.00000      37  9.2   65     6  18 2.3223877
```



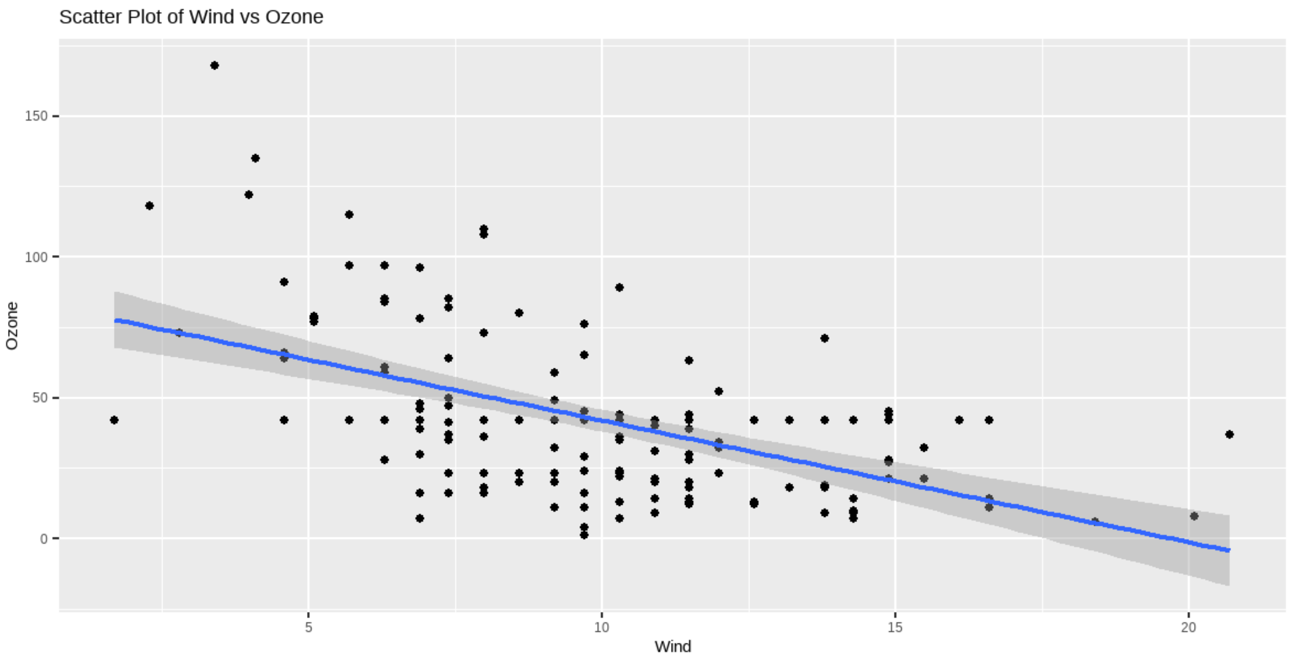
```
> # Log transformation of Wind (adding 1 to avoid log(0))
> airquality$Log_Wind <- log(airquality$Wind + 1)
> print("Log transformed Wind column:")
[1] "Log transformed Wind column:"
> print(head(airquality$Log_Wind, 10))
 [1] 2.128232 2.197225 2.610070 2.525729 2.727853 2.766319 2.261763 2.694627 3.049273 2.261763
> 
> print("21BDS0085 JVNGANESH")
[1] "21BDS0085 JVNGANESH"
> # 3. Imputation of Missing Values
> # Mean Imputation for Ozone
> airquality$Ozone[is.na(airquality$Ozone)] <- mean(airquality$Ozone, na.rm = TRUE)
> print("After Mean Imputation of Ozone column:")
[1] "After Mean Imputation of Ozone column:"
> print(summary(airquality$Ozone))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00   21.00   42.13   42.13   46.00  168.00 
> 
> # Median Imputation for Solar.R
> airquality$Solar.R[is.na(airquality$Solar.R)] <- median(airquality$Solar.R, na.rm = TRUE)
> print("After Median Imputation of Solar.R column:")
[1] "After Median Imputation of Solar.R column:"
> print(summary(airquality$Solar.R))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    7.0   120.0   205.0   186.8   256.0   334.0 
> 
> # Class-based Imputation (Imputation by median within groups)
> airquality <- airquality %>%
+   group_by(Month) %>%
+   mutate(Ozone = ifelse(is.na(Ozone), median(Ozone, na.rm = TRUE), Ozone))
> print("Class-based imputation of Ozone column by Month:")
[1] "Class-based imputation of Ozone column by Month:"
> print(airquality)
# A tibble: 153 × 8
# Groups:   Month [5]
   Ozone Solar.R  Wind  Temp Month   Day Ozone_scaled Log_Wind
   <dbl>   <dbl> <dbl> <int> <int> <int>        <dbl>    <dbl>
 1  41       190   7.4    67     5     1      -0.0342     2.13
 2  36       118   8      72     5     2      -0.186      2.20
 3  12       149  12.6    74     5     3      -0.913      2.61
 4  18       313  11.5    62     5     4      -0.731      2.53
 5  42.1     205  14.3    56     5     5      NA          2.73
 6  28       205  14.9    66     5     6      -0.428      2.77
 7  23       299   8.6    65     5     7      -0.580      2.26
 8  19        99  13.8    59     5     8      -0.701      2.69
 9   8        19  20.1    61     5     9      -1.03       3.05
10  42.1     194   8.6    69     5    10      NA          2.26
# ℹ 143 more rows
# ℹ Use `print(n = ...)` to see more rows
> 
> print("21BDS0085 JVNGANESH")
[1] "21BDS0085 JVNGANESH"
> 
```

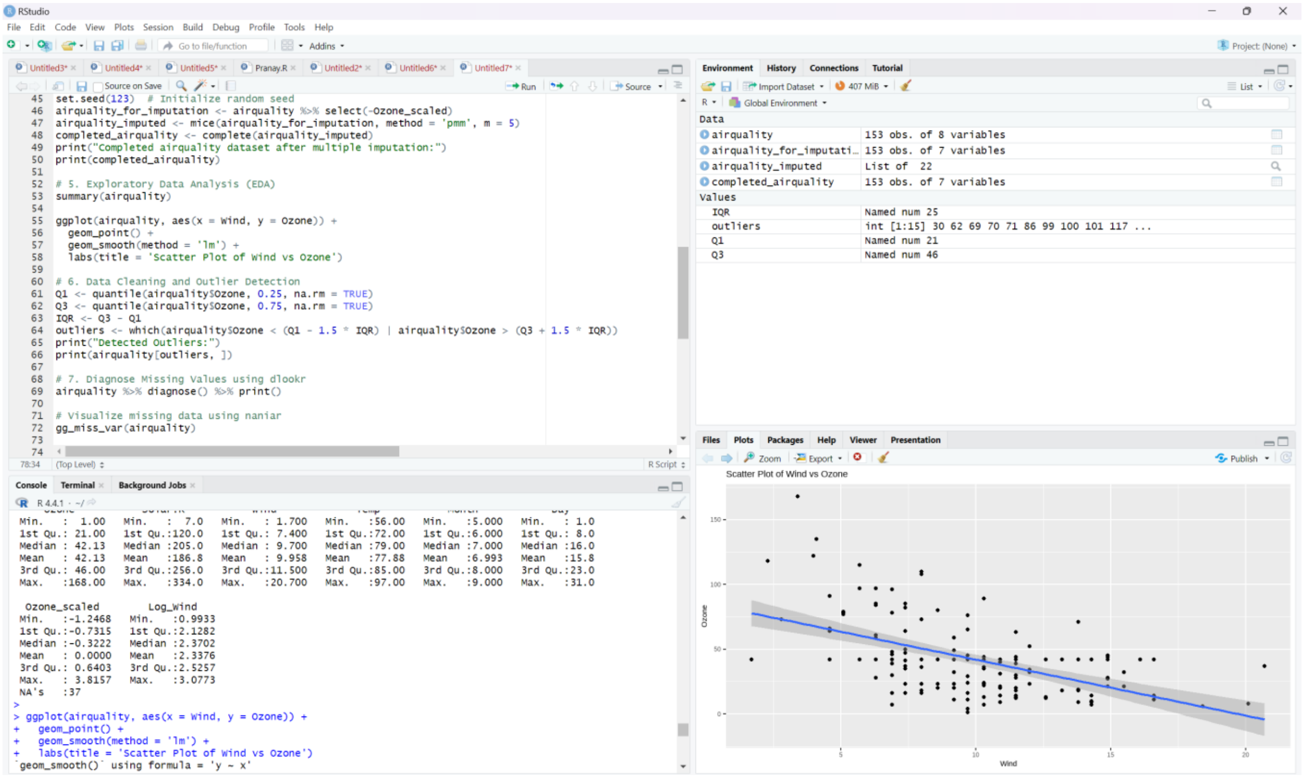


```
75   42.12931      291 14.9     91       7  14 2.7663191
76    7.00000       48 14.3     80       7  15 2.7278528
77   48.00000      260  6.9     81       7  16 2.0668628
78   35.00000      274 10.3     82       7  17 2.4248027
79   61.00000      285  6.3     84       7  18 1.9878743
80   79.00000      187  5.1     87       7  19 1.8082888
81   63.00000      220 11.5     85       7  20 2.5257286
82   16.00000        7  6.9     74       7  21 2.0668628
83   42.12931      258  9.7     81       7  22 2.3702437
84   42.12931      295 11.5     82       7  23 2.5257286
85   80.00000      294  8.6     86       7  24 2.2617631
86  108.00000      223  8.0     85       7  25 2.1972246
87   20.00000       81  8.6     82       7  26 2.2617631
88   52.00000       82 12.0     86       7  27 2.5649494
89   82.00000      213  7.4     88       7  28 2.1282317
90   50.00000      275  7.4     86       7  29 2.1282317
91   64.00000      253  7.4     83       7  30 2.1282317
92   59.00000      254  9.2     81       7  31 2.3223877
93   39.00000       83  6.9     81       8   1 2.0668628
94    9.00000       24 13.8     81       8   2 2.6946272
95   16.00000       77  7.4     82       8   3 2.1282317
96   78.00000      205  6.9     86       8   4 2.0668628
97   35.00000      205  7.4     85       8   5 2.1282317
98   66.00000      205  4.6     87       8   6 1.7227666
99  122.00000      255  4.0     89       8   7 1.6094379
100  89.00000      229 10.3     90       8   8 2.4248027
101 110.00000      207  8.0     90       8   9 2.1972246
102  42.12931      222  8.6     92       8  10 2.2617631
103  42.12931      137 11.5     86       8  11 2.5257286
104  44.00000      192 11.5     86       8  12 2.5257286
105  28.00000      273 11.5     82       8  13 2.5257286
106  65.00000      157  9.7     80       8  14 2.3702437
107  42.12931       64 11.5     79       8  15 2.5257286
108  22.00000       71 10.3     77       8  16 2.4248027
109  59.00000       51  6.3     79       8  17 1.9878743
110  23.00000      115  7.4     76       8  18 2.1282317
111  31.00000      244 10.9     78       8  19 2.4765384
112  44.00000      190 10.3     78       8  20 2.4248027
113  21.00000      259 15.5     77       8  21 2.8033604
114   9.00000       36 14.3     72       8  22 2.7278528
115  42.12931      255 12.6     75       8  23 2.6100698
116  45.00000      212  9.7     79       8  24 2.3702437
117 168.00000      238  3.4     81       8  25 1.4816045
118  73.00000      215  8.0     86       8  26 2.1972246
119  42.12931      153  5.7     88       8  27 1.9021075
120  76.00000      203  9.7     97       8  28 2.3702437
121 118.00000      225  2.3     94       8  29 1.1939225
122  84.00000      237  6.3     96       8  30 1.9878743
123  85.00000      188  6.3     94       8  31 1.9878743
124  96.00000      167  6.9     91       9   1 2.0668628
125  78.00000      197  5.1     92       9   2 1.8082888
126  73.00000      183  2.8     93       9   3 1.3350011
127  91.00000      189  4.6     93       9   4 1.7227666
128  47.00000       95  7.4     87       9   5 2.1282317
```



```
116  45.00000    212  9.7   79     8  24 2.3702437
117 168.00000    238  3.4   81     8  25 1.4816045
118  73.00000    215  8.0   86     8  26 2.1972246
119  42.12931    153  5.7   88     8  27 1.9021075
120  76.00000    203  9.7   97     8  28 2.3702437
121 118.00000    225  2.3   94     8  29 1.1939225
122  84.00000    237  6.3   96     8  30 1.9878743
123  85.00000    188  6.3   94     8  31 1.9878743
124  96.00000    167  6.9   91     9   1 2.0668628
125  78.00000    197  5.1   92     9   2 1.8082888
126  73.00000    183  2.8   93     9   3 1.3350011
127  91.00000    189  4.6   93     9   4 1.7227666
128  47.00000     95  7.4   87     9   5 2.1282317
129  32.00000     92 15.5   84     9   6 2.8033604
130  20.00000    252 10.9   80     9   7 2.4765384
131  23.00000    220 10.3   78     9   8 2.4248027
132  21.00000    230 10.9   75     9   9 2.4765384
133  24.00000    259  9.7   73     9  10 2.3702437
134  44.00000    236 14.9   81     9  11 2.7663191
135  21.00000    259 15.5   76     9  12 2.8033604
136  28.00000    238  6.3   77     9  13 1.9878743
137   9.00000     24 10.9   71     9  14 2.4765384
138  13.00000    112 11.5   71     9  15 2.5257286
139  46.00000    237  6.9   78     9  16 2.0668628
140  18.00000    224 13.8   67     9  17 2.6946272
141  13.00000     27 10.3   76     9  18 2.4248027
142  24.00000    238 10.3   68     9  19 2.4248027
 [ reached 'max' / getOption("max.print") -- omitted 11 rows ]
> # 5. Exploratory Data Analysis (EDA)
> summary(airquality)
     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 : 42.13   Median :205.0   Median : 9.700   Median :79.00   Median :7.000   Median :16.0  
 Mean   : 42.13   Mean   :186.8   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  
                                                                                                 
  Ozone_scaled        Log_Wind     
 Min.   :-1.2468   Min.   :0.9933  
 1st Qu.:-0.7315   1st Qu.:2.1282  
 Median :-0.3222   Median :2.3702  
 Mean   : 0.0000   Mean   :2.3376  
 3rd Qu.: 0.6403   3rd Qu.:2.5257  
 Max.   : 3.8157   Max.   :3.0773  
 NA's   :37                        
> 
> ggplot(airquality, aes(x = Wind, y = Ozone)) +
+   geom_point() +
+   geom_smooth(method = 'lm') +
+   labs(title = 'Scatter Plot of Wind vs Ozone')
`geom_smooth()` using formula = 'y ~ x'
> 
```

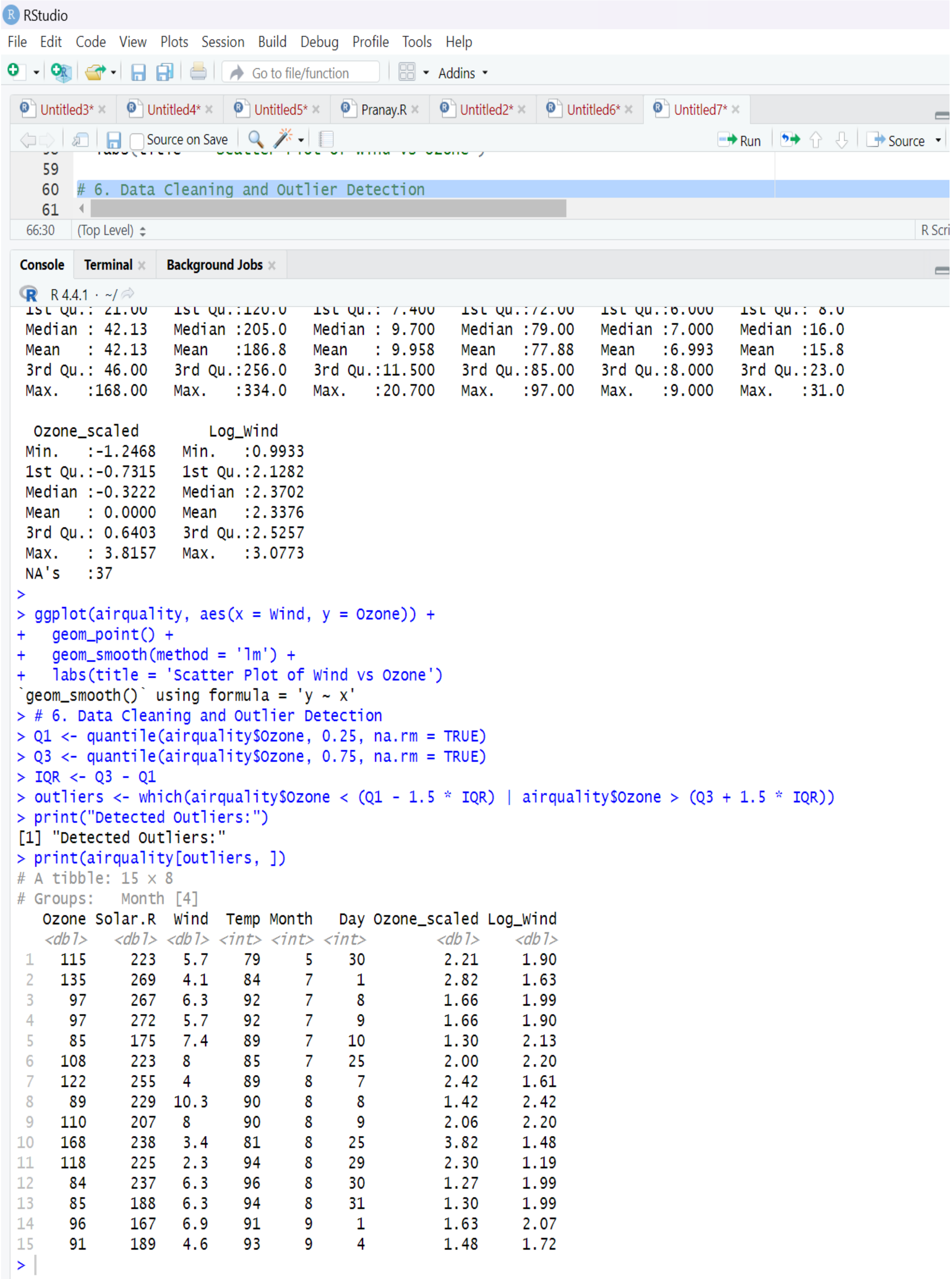




```r
set.seed(123)  # Initialize random seed
airquality_for_imputation <- airquality %>% select(-Ozone_scaled)
airquality_imputed <- mice(airquality_for_imputation, method = 'pmm', m = 5)
completed_airquality <- complete(airquality_imputed)
print("Completed airquality dataset after multiple imputation:")
print(completed_airquality)

# 5. Exploratory Data Analysis (EDA)
summary(airquality)

ggplot(airquality, aes(x = Wind, y = Ozone)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  labs(title = 'Scatter Plot of Wind vs Ozone')

# 6. Data Cleaning and Outlier Detection
Q1 <- quantile(airquality$Ozone, 0.25, na.rm = TRUE)
Q3 <- quantile(airquality$Ozone, 0.75, na.rm = TRUE)
IQR <- Q3 - Q1
outliers <- which(airquality$Ozone < (Q1 - 1.5 * IQR) | airquality$Ozone > (Q3 + 1.5 * IQR))
print("Detected Outliers:")
print(airquality[outliers, ])

# 7. Diagnose Missing Values using dlookr
airquality %>% diagnose() %>% print()

# Visualize missing data using naniar
gg_miss_var(airquality)

```

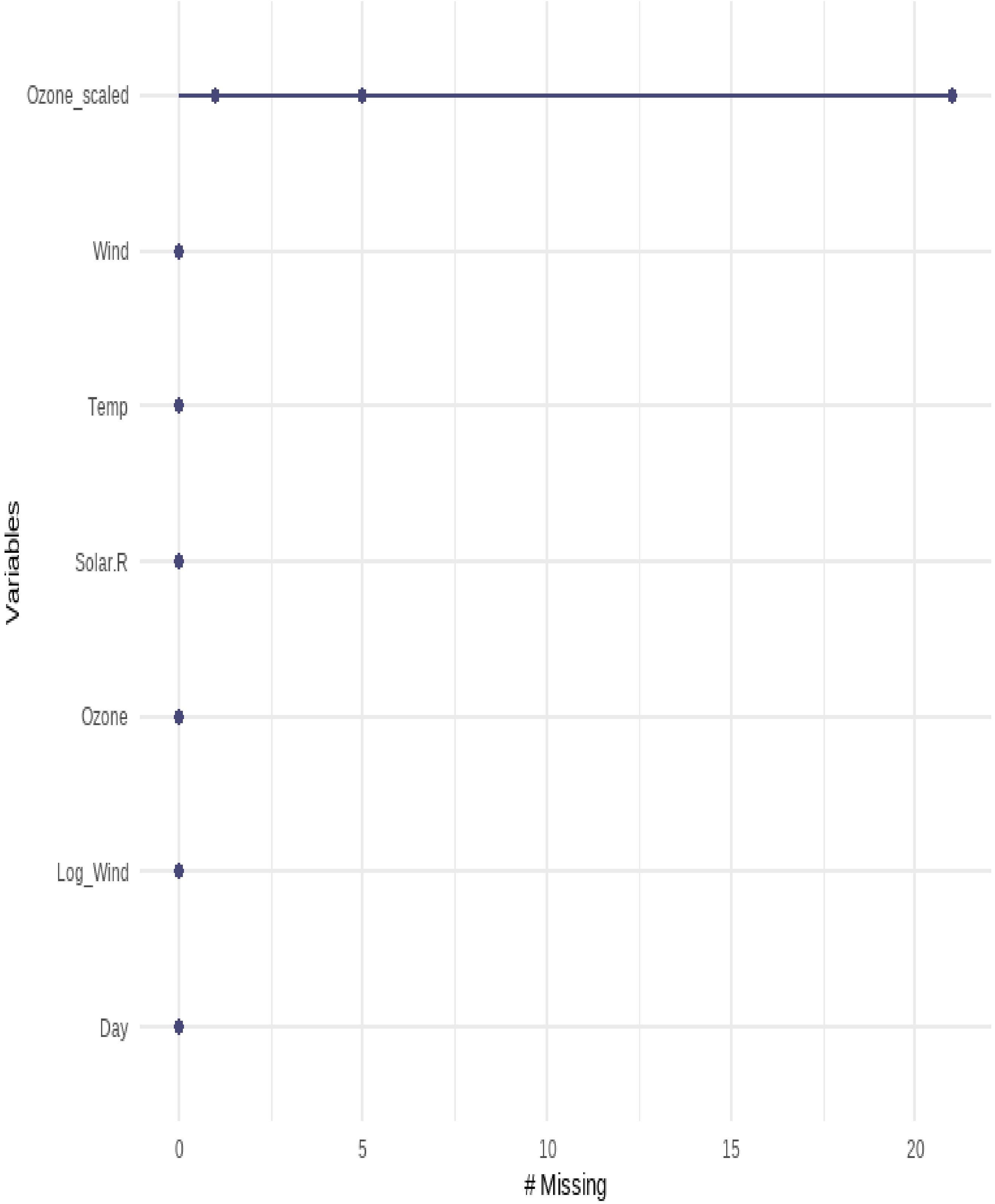
```
Data
airquality                     153 obs. of 8 variables
airquality_for_imputati…       153 obs. of 7 variables
airquality_imputed             List of  22
completed_airquality           153 obs. of 7 variables
Values
  IQR                          Named num 25
  outliers                     int [1:15] 30 62 69 70 71 86 99 100 101 117 ...
  Q1                           Named num 21
  Q3                           Named num 46
```

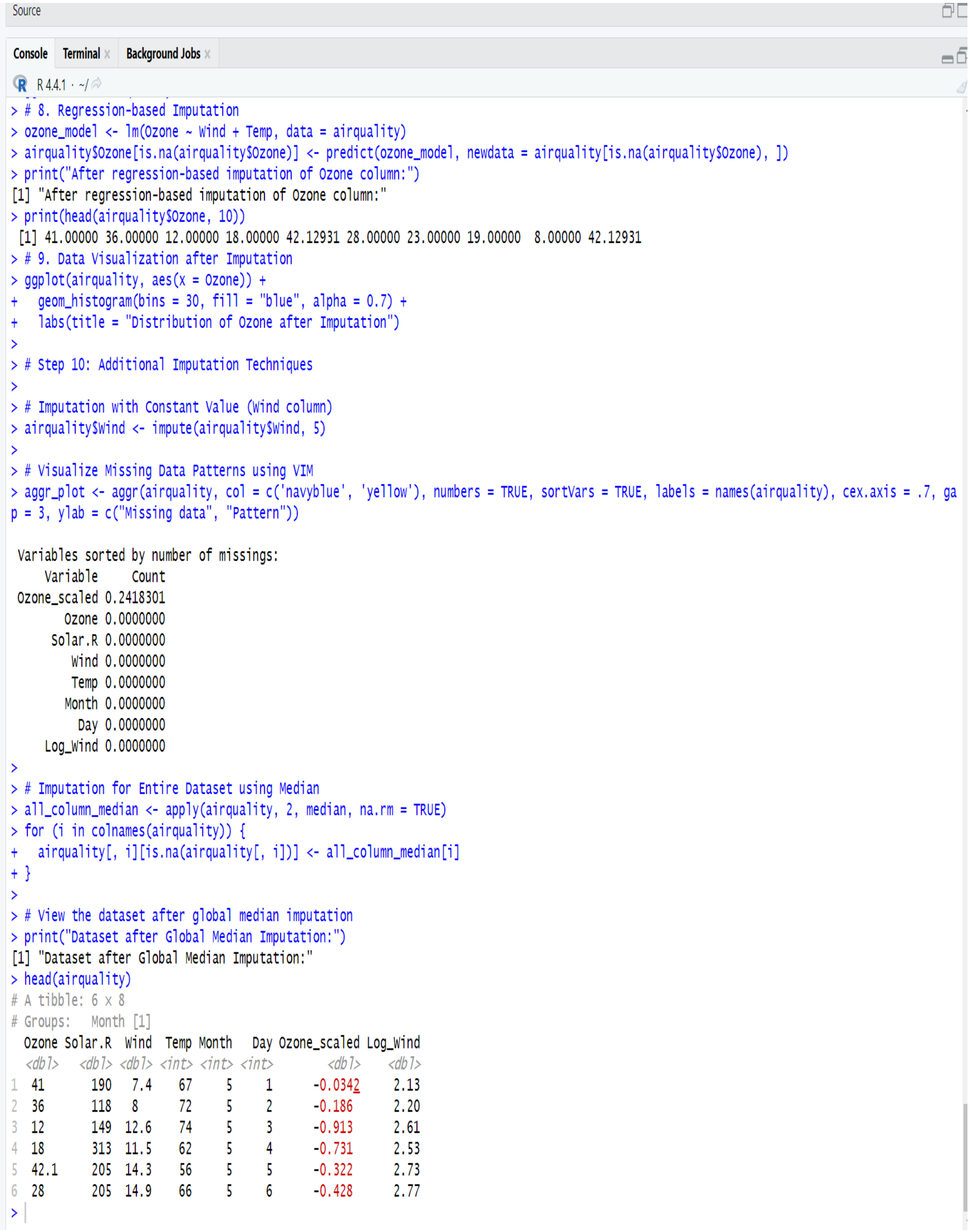
```
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00   Min.   :5.000   Min.   : 1.0
 1st Qu.: 21.00   1st Qu.:120.0   1st Qu.: 7.400   1st Qu.:72.00   1st Qu.:6.000   1st Qu.: 8.0
 Median : 42.13   Median :205.0   Median : 9.700   Median :79.00   Median :7.000   Median :16.0
 Mean   : 42.13   Mean   :186.8   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

  Ozone_scaled        Log_Wind
 Min.   :-1.2468   Min.   :0.9933
 1st Qu.:-0.7315   1st Qu.:2.1282
 Median :-0.3222   Median :2.3702
 Mean   : 0.0000   Mean   :2.3376
 3rd Qu.: 0.6403   3rd Qu.:2.5257
 Max.   : 3.8157   Max.   :3.0773
 NA's   :37
>
> ggplot(airquality, aes(x = Wind, y = Ozone)) +
+   geom_point() +
+   geom_smooth(method = 'lm') +
+   labs(title = 'Scatter Plot of Wind vs Ozone')
`geom_smooth()` using formula = 'y ~ x'
```


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Untitled3* Untitled4* Untitled5* Pranay.R Untitled2* Untitled6* Untitled7*
Source on Save
Run
Source
66:30
(Top Level)
R Scri
Console
Terminal
Background Jobs
R 4.4.1 · ~/

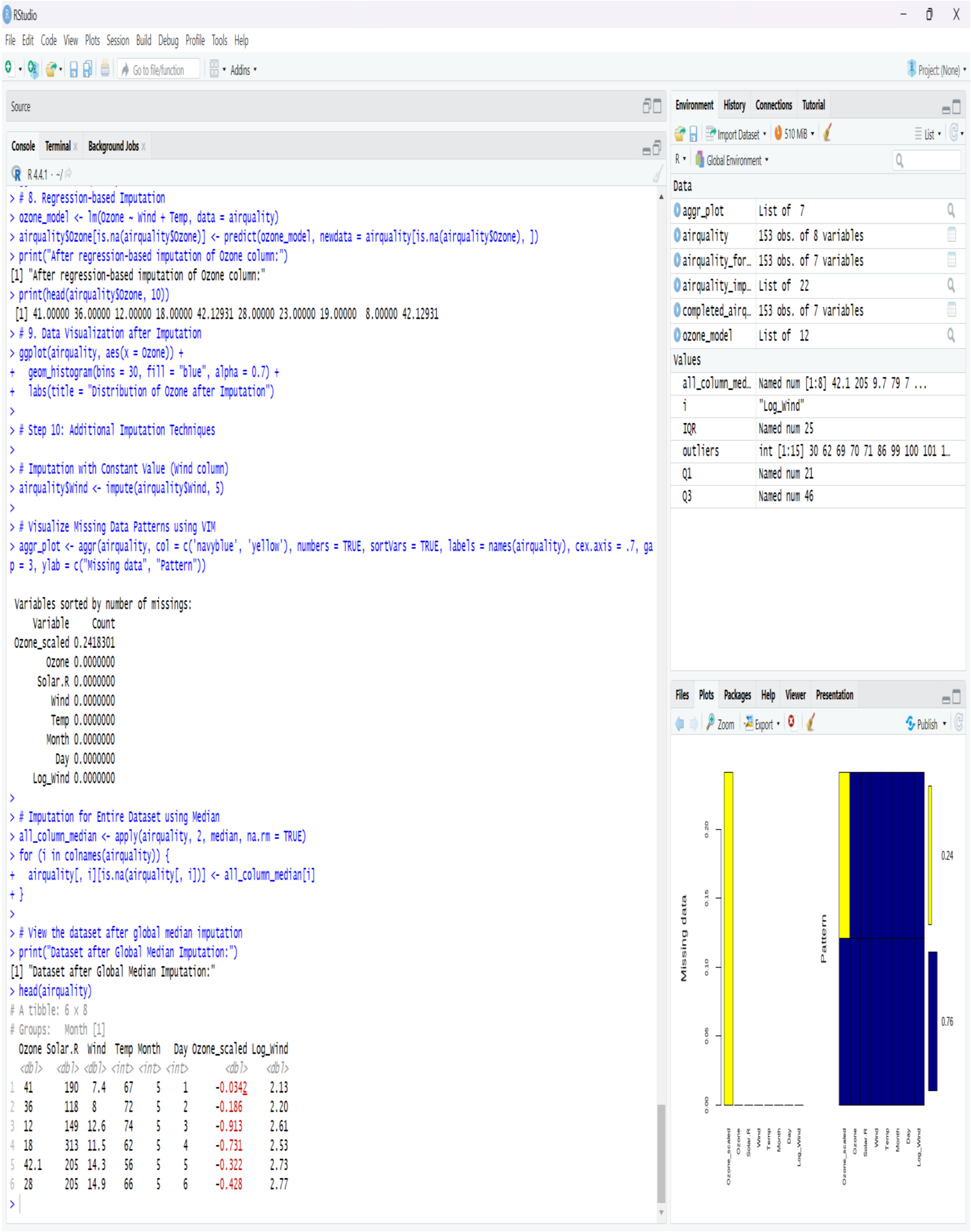
```
  labs(title = 'Scatter Plot of Wind vs Ozone')

# 6. Data Cleaning and Outlier Detection
```

```
 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 : 42.13  Median :205.0  Median : 9.700  Median :79.00  Median :7.000  Median :16.0
 Mean   : 42.13  Mean   :186.8  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

  Ozone_scaled        Log_Wind
 Min.   :-1.2468   Min.   :0.9933
 1st Qu.:-0.7315   1st Qu.:2.1282
 Median :-0.3222   Median :2.3702
 Mean   : 0.0000   Mean   :2.3376
 3rd Qu.: 0.6403   3rd Qu.:2.5257
 Max.   : 3.8157   Max.   :3.0773
 NA's   :37
>
> ggplot(airquality, aes(x = Wind, y = Ozone)) +
+   geom_point() +
+   geom_smooth(method = 'lm') +
+   labs(title = 'Scatter Plot of Wind vs Ozone')
`geom_smooth()` using formula = 'y ~ x'
> # 6. Data Cleaning and Outlier Detection
> Q1 <- quantile(airquality$Ozone, 0.25, na.rm = TRUE)
> Q3 <- quantile(airquality$Ozone, 0.75, na.rm = TRUE)
> IQR <- Q3 - Q1
> outliers <- which(airquality$Ozone < (Q1 - 1.5 * IQR) | airquality$Ozone > (Q3 + 1.5 * IQR))
> print("Detected Outliers:")
[1] "Detected Outliers:"
> print(airquality[outliers, ])
# A tibble: 15 × 8
# Groups:   Month [4]
   Ozone Solar.R  Wind  Temp Month   Day Ozone_scaled Log_Wind
   <dbl>   <dbl> <dbl> <int> <int> <int>        <dbl>    <dbl>
 1   115     223   5.7    79     5    30         2.21     1.90
 2   135     269   4.1    84     7     1         2.82     1.63
 3    97     267   6.3    92     7     8         1.66     1.99
 4    97     272   5.7    92     7     9         1.66     1.90
 5    85     175   7.4    89     7    10         1.30     2.13
 6   108     223   8      85     7    25         2.00     2.20
 7   122     255   4      89     8     7         2.42     1.61
 8    89     229  10.3    90     8     8         1.42     2.42
 9   110     207   8      90     8     9         2.06     2.20
10   168     238   3.4    81     8    25         3.82     1.48
11   118     225   2.3    94     8    29         2.30     1.19
12    84     237   6.3    96     8    30         1.27     1.99
13    85     188   6.3    94     8    31         1.30     1.99
14    96     167   6.9    91     9     1         1.63     2.07
15    91     189   4.6    93     9     4         1.48     1.72
> 
```


Ozone_scaled
Wind
Temp
Solar.R
Ozone
Log_Wind
Day
Variables
0
5
10
15
20
# Missing



```
> # 8. Regression-based Imputation
> ozone_model <- lm(Ozone ~ Wind + Temp, data = airquality)
> airquality$Ozone[is.na(airquality$Ozone)] <- predict(ozone_model, newdata = airquality[is.na(airquality$Ozone), ])
> print("After regression-based imputation of Ozone column:")
[1] "After regression-based imputation of Ozone column:"
> print(head(airquality$Ozone, 10))
 [1] 41.00000 36.00000 12.00000 18.00000 42.12931 28.00000 23.00000 19.00000  8.00000 42.12931
> # 9. Data Visualization after Imputation
> ggplot(airquality, aes(x = Ozone)) +
+   geom_histogram(bins = 30, fill = "blue", alpha = 0.7) +
+   labs(title = "Distribution of Ozone after Imputation")
> 
> # Step 10: Additional Imputation Techniques
> 
> # Imputation with Constant Value (Wind column)
> airquality$Wind <- impute(airquality$Wind, 5)
> 
> # Visualize Missing Data Patterns using VIM
> aggr_plot <- aggr(airquality, col = c('navyblue', 'yellow'), numbers = TRUE, sortVars = TRUE, labels = names(airquality), cex.axis = .7, ga
p = 3, ylab = c("Missing data", "Pattern"))

 Variables sorted by number of missings:
     Variable     Count
 Ozone_scaled 0.2418301
        Ozone 0.0000000
      Solar.R 0.0000000
         Wind 0.0000000
         Temp 0.0000000
        Month 0.0000000
          Day 0.0000000
     Log_Wind 0.0000000
> 
> # Imputation for Entire Dataset using Median
> all_column_median <- apply(airquality, 2, median, na.rm = TRUE)
> for (i in colnames(airquality)) {
+   airquality[, i][is.na(airquality[, i])] <- all_column_median[i]
+ }
> 
> # View the dataset after global median imputation
> print("Dataset after Global Median Imputation:")
[1] "Dataset after Global Median Imputation:"
> head(airquality)
# A tibble: 6 x 8
# Groups:   Month [1]
  Ozone Solar.R  Wind  Temp Month   Day Ozone_scaled Log_Wind
  <dbl>   <dbl> <dbl> <int> <int> <int>        <dbl>    <dbl>
1  41       190   7.4    67     5     1      -0.0342     2.13
2  36       118   8      72     5     2      -0.186      2.20
3  12       149  12.6    74     5     3      -0.913      2.61
4  18       313  11.5    62     5     4      -0.731      2.53
5  42.1     205  14.3    56     5     5      -0.322      2.73
6  28       205  14.9    66     5     6      -0.428      2.77
> 
```



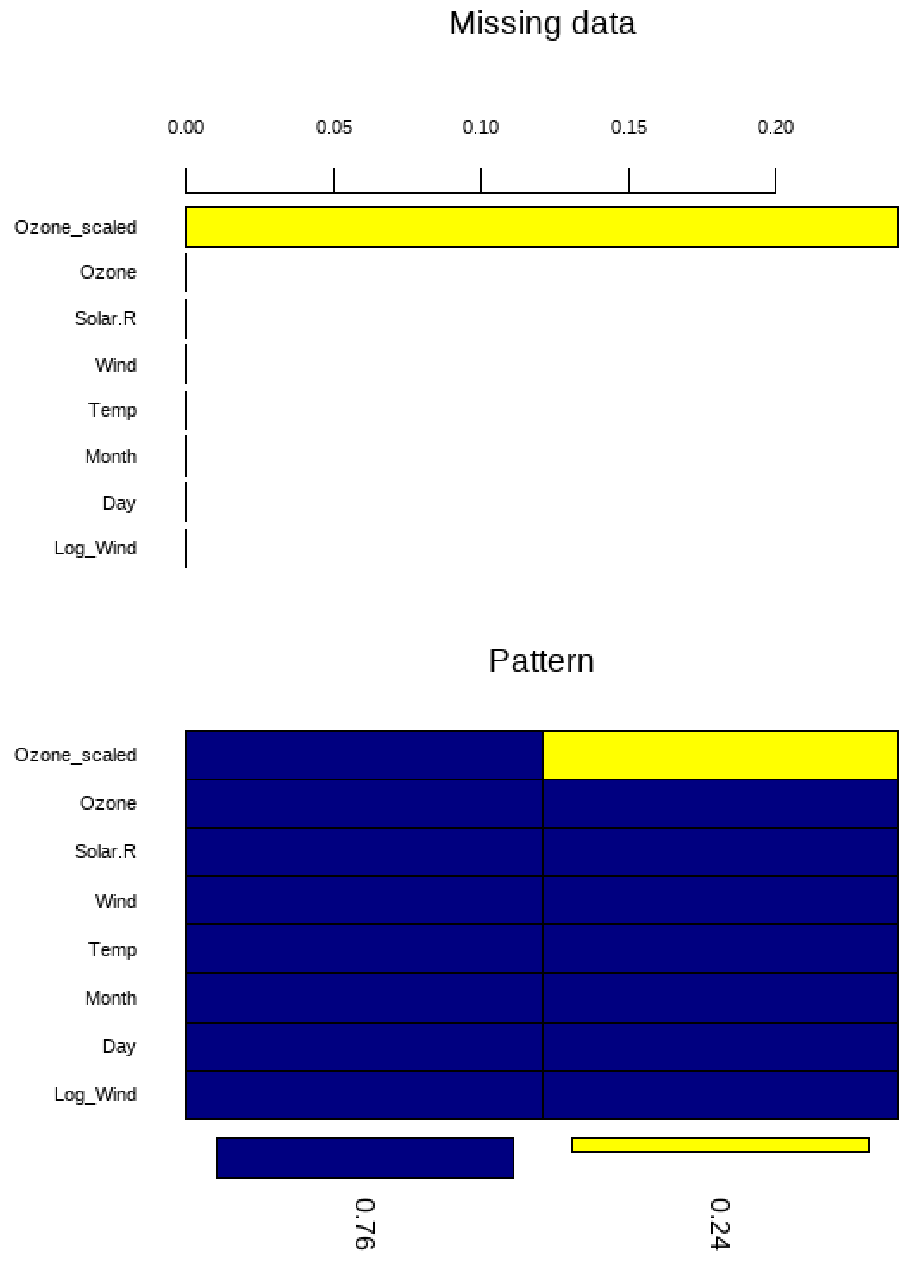
```
> # 8. Regression-based Imputation
> ozone_model <- lm(Ozone ~ Wind + Temp, data = airquality)
> airquality$Ozone[is.na(airquality$Ozone)] <- predict(ozone_model, newdata = airquality[is.na(airquality$Ozone), ])
> print("After regression-based imputation of Ozone column:")
[1] "After regression-based imputation of Ozone column:"
> print(head(airquality$Ozone, 10))
 [1] 41.00000 36.00000 12.00000 18.00000 42.12931 28.00000 23.00000 19.00000  8.00000 42.12931
> # 9. Data Visualization after Imputation
> ggplot(airquality, aes(x = Ozone)) +
+   geom_histogram(bins = 30, fill = "blue", alpha = 0.7) +
+   labs(title = "Distribution of Ozone after Imputation")
>
> # Step 10: Additional Imputation Techniques
>
> # Imputation with Constant Value (Wind column)
> airquality$Wind <- impute(airquality$Wind, 5)
>
> # Visualize Missing Data Patterns using VIM
> aggr_plot <- aggr(airquality, col = c('navyblue', 'yellow'), numbers = TRUE, sortVars = TRUE, labels = names(airquality), cex.axis = .7, ga
p = 3, ylab = c("Missing data", "Pattern"))

 Variables sorted by number of missings:
     Variable     Count
 Ozone_scaled 0.2418301
        Ozone 0.0000000
      Solar.R 0.0000000
         Wind 0.0000000
         Temp 0.0000000
        Month 0.0000000
          Day 0.0000000
     Log_Wind 0.0000000
>
> # Imputation for Entire Dataset using Median
> all_column_median <- apply(airquality, 2, median, na.rm = TRUE)
> for (i in colnames(airquality)) {
+   airquality[, i][is.na(airquality[, i])] <- all_column_median[i]
+ }
>
> # View the dataset after global median imputation
> print("Dataset after Global Median Imputation:")
[1] "Dataset after Global Median Imputation:"
> head(airquality)
# A tibble: 6 x 8
# Groups:   Month [1]
  Ozone Solar.R  Wind  Temp Month   Day Ozone_scaled Log_Wind
  <dbl>   <dbl> <dbl> <int> <int> <int>        <dbl>    <dbl>
1  41       190   7.4    67     5     1      -0.0342     2.13
2  36       118   8      72     5     2      -0.186      2.20
3  12       149  12.6    74     5     3      -0.913      2.61
4  18       313  11.5    62     5     4      -0.731      2.53
5  42.1     205  14.3    56     5     5      -0.322      2.73
6  28       205  14.9    66     5     6      -0.428      2.77
>
```

Data

aggr_plot	List of 7
airquality	153 obs. of 8 variables
airquality_for…	153 obs. of 7 variables
airquality_imp…	List of 22
completed_airq…	153 obs. of 7 variables
ozone_model	List of 12

Values

all_column_med…	Named num [1:8] 42.1 205 9.7 79 7 ...
i	"Log_Wind"
IQR	Named num 25
outliers	int [1:15] 30 62 69 70 71 86 99 100 101 1…
Q1	Named num 21
Q3	Named num 46


Missing data
0.00
0.05
0.10
0.15
0.20
Ozone_scaled
Ozone
Solar.R
Wind
Temp
Month
Day
Log_Wind
Pattern
Ozone_scaled
Ozone
Solar.R
Wind
Temp
Month
Day
Log_Wind
0.76
0.24

CODE :

```
print("21BDS0085 JVNGANESH")

#

# #Install necessary libraries (uncomment if not already installed)

# install.packages("Hmisc")    # For single variable imputation

# install.packages("mice")     # For multiple imputation methods

# install.packages("VIM")      # For visualizing missing data

# install.packages("dplyr")    # For data manipulation




# Load necessary libraries

library(dlookr)

library(dplyr)

library(ggplot2)

library(mice)

library(naniar)  # For visualizing missing data

library(Hmisc)   # For single variable imputation

library(VIM)     # For missing data visualization


# Load default dataset - airquality

data("airquality")

print("21BDS0085 JVNGANESH")


# 1. View basic summary

print("Basic Summary of airquality dataset:")

summary(airquality)
```

```
# 2. Data Transformation (Standardization and Log Transformation)

airquality$Ozone_scaled <- as.numeric(scale(airquality$Ozone, center = TRUE, scale =
TRUE))

print("Standardized Ozone column:")

print(head(airquality$Ozone_scaled, 10))


airquality$Log_Wind <- log(airquality$Wind + 1)

print("Log transformed Wind column:")

print(head(airquality$Log_Wind, 10))


# 3. Imputation of Missing Values

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

airquality$Solar.R[is.na(airquality$Solar.R)] <- median(airquality$Solar.R, na.rm = TRUE)


airquality <- airquality %>%

  group_by(Month) %>%

  mutate(Ozone = ifelse(is.na(Ozone), median(Ozone, na.rm = TRUE), Ozone))


# 4. Multiple Imputation Using MICE

set.seed(123)  # Initialize random seed

airquality_for_imputation <- airquality %>% select(-Ozone_scaled)

airquality_imputed <- mice(airquality_for_imputation, method = 'pmm', m = 5)

completed_airquality <- complete(airquality_imputed)

print("Completed airquality dataset after multiple imputation:")

print(completed_airquality)
```

```
# 5. Exploratory Data Analysis (EDA)
summary(airquality)


ggplot(airquality, aes(x = Wind, y = Ozone)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  labs(title = 'Scatter Plot of Wind vs Ozone')


# 6. Data Cleaning and Outlier Detection
Q1 <- quantile(airquality$Ozone, 0.25, na.rm = TRUE)
Q3 <- quantile(airquality$Ozone, 0.75, na.rm = TRUE)
IQR <- Q3 - Q1
outliers <- which(airquality$Ozone < (Q1 - 1.5 * IQR) | airquality$Ozone > (Q3 + 1.5 * IQR))
print("Detected Outliers:")
print(airquality[outliers, ])


# 7. Diagnose Missing Values using dlookr
airquality %>% diagnose() %>% print()


# Visualize missing data using naniar
gg_miss_var(airquality)


# 8. Regression-based Imputation
ozone_model <- lm(Ozone ~ Wind + Temp, data = airquality)
```

```
airquality$Ozone[is.na(airquality$Ozone)] <- predict(ozone_model, newdata =
airquality[is.na(airquality$Ozone), ])
print("After regression-based imputation of Ozone column:")
print(head(airquality$Ozone, 10))

# 9. Data Visualization after Imputation
ggplot(airquality, aes(x = Ozone)) +
  geom_histogram(bins = 30, fill = "blue", alpha = 0.7) +
  labs(title = "Distribution of Ozone after Imputation")

# Step 10: Additional Imputation Techniques

# Imputation with Constant Value (Wind column)
airquality$Wind <- impute(airquality$Wind, 5)

# Visualize Missing Data Patterns using VIM
aggr_plot <- aggr(airquality, col = c('navyblue', 'yellow'), numbers = TRUE, sortVars = TRUE,
labels = names(airquality), cex.axis = .7, gap = 3, ylab = c("Missing data", "Pattern"))

# Imputation for Entire Dataset using Median
all_column_median <- apply(airquality, 2, median, na.rm = TRUE)
for (i in colnames(airquality)) {
  airquality[, i][is.na(airquality[, i])] <- all_column_median[i]
}

# View the dataset after global median imputation
print("Dataset after Global Median Imputation:")
```

head(airquality)

OUTPUTS:

```
 > install.packages("naniar")
 Error in install.packages : Updating loaded packages
 > library(naniar)
 >

 Restarting R session...

 > install.packages("naniar")
 WARNING: Rtools is required to build R packages but is not currently installed. Please download and inst

 https://cran.rstudio.com/bin/windows/Rtools/
 Installing package into ‘C:/Users/lenovo/AppData/Local/R/win-library/4.4’
 (as ‘lib’ is unspecified)
 trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.4/naniar_1.1.0.zip'
 Content type 'application/zip' length 2766333 bytes (2.6 MB)
 downloaded 2.6 MB

 package ‘naniar’ successfully unpacked and MD5 sums checked

 The downloaded binary packages are in
         C:\Users\lenovo\AppData\Local\Temp\RtmpiSdJTM\downloaded_packages
 > print("21BDS0085 JVNGANESH")
 [1] "21BDS0085 JVNGANESH"
```

```
> # Load necessary libraries
> library(dlookr)
Registered S3 methods overwritten by 'dlookr':
  method         from
  plot.transform  scales
  print.transform scales
Because it is an offline environment, only offline fonts are imported.

Attaching package: ‘dlookr’

The following object is masked from ‘package:base’:

    transform

Warning message:
In file.rename(tmp, destfile) :
  cannot rename file 'C:\Users\lenovo\AppData\Local\Temp\RtmpiSdJTM\PbykFmXiEBPT4ITbgNA5Cgm20
'C:\Users\lenovo\AppData\Local\Temp\RtmpiSdJTM\PbykFmXiEBPT4ITbgNA5Cgm20HTs4JMMuA.otf', rea
> library(dplyr)

Attaching package: ‘dplyr’

The following objects are masked from ‘package:stats’:

    filter, lag
```

```
The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union

> library(ggplot2)
> library(mice)

Attaching package: ‘mice’

The following object is masked from ‘package:stats’:

    filter

The following objects are masked from ‘package:base’:

    cbind, rbind

> library(naniar)  # For visualizing missing data
> library(Hmisc)   # For single variable imputation

Attaching package: ‘Hmisc’

The following objects are masked from ‘package:dplyr’:

    src, summarize
```

```
The following object is masked from ‘package:dlookr’:

    describe

The following objects are masked from ‘package:base’:

    format.pval, units

> library(VIM)     # For missing data visualization
Error in library(VIM) : there is no package called ‘VIM’
> Install necessary libraries (uncomment if not already installed)
Error: unexpected symbol in "Install necessary"
> #Install necessary libraries (uncomment if not already installed)
> install.packages("Hmisc")     # For single variable imputation
Error in install.packages : Updating loaded packages
> install.packages("mice")      # For multiple imputation methods
Error in install.packages : Updating loaded packages
> install.packages("VIM")       # For visualizing missing data
Error in install.packages : Updating loaded packages
> install.packages("dplyr")     # For data manipulation
Error in install.packages : Updating loaded packages

Restarting R session...
```

```
> install.packages("Hmisc")

WARNING: Rtools is required to build R packages but is not currently installed. Please download and inst


https://cran.rstudio.com/bin/windows/Rtools/

Installing package into ‘C:/Users/lenovo/AppData/Local/R/win-library/4.4’

(as ‘lib’ is unspecified)

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.4/Hmisc_5.1-3.zip'

Content type 'application/zip' length 3606271 bytes (3.4 MB)

downloaded 3.4 MB


package ‘Hmisc’ successfully unpacked and MD5 sums checked


The downloaded binary packages are in

        C:\Users\lenovo\AppData\Local\Temp\RtmpaOWN9Z\downloaded_packages


Restarting R session...


> install.packages("mice")

WARNING: Rtools is required to build R packages but is not currently installed. Please download and inst


https://cran.rstudio.com/bin/windows/Rtools/

Installing package into ‘C:/Users/lenovo/AppData/Local/R/win-library/4.4’

(as ‘lib’ is unspecified)

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.4/mice_3.16.0.zip'

Content type 'application/zip' length 1882211 bytes (1.8 MB)
```

```
downloaded 1.8 MB

package ‘mice’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\lenovo\AppData\Local\Temp\Rtmpob3omE\downloaded_packages
> library(dlookr)
Registered S3 methods overwritten by 'dlookr':
  method         from  
  plot.transform  scales
  print.transform scales

Attaching package: ‘dlookr’

The following object is masked from ‘package:base’:

    transform

> library(dplyr)

Attaching package: ‘dplyr’

The following objects are masked from ‘package:stats’:

    filter, lag
```

```
The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union

> library(ggplot2)
> library(mice)

Attaching package: ‘mice’

The following object is masked from ‘package:stats’:

    filter

The following objects are masked from ‘package:base’:

    cbind, rbind

> library(naniar)  # For visualizing missing data
> library(Hmisc)   # For single variable imputation

Attaching package: ‘Hmisc’

The following objects are masked from ‘package:dplyr’:
    src, summarize


The following object is masked from ‘package:dlookr’:


    describe


The following objects are masked from ‘package:base’:


    format.pval, units


> library(VIM)     # For missing data visualization
Loading required package: colorspace
Loading required package: grid
VIM is ready to use.


Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues


Attaching package: ‘VIM’


The following object is masked from ‘package:datasets’:


    sleep


>
> # Load default dataset - airquality
> data("airquality")
> print("21BDS0085 JVNGANESH")
[1] "21BDS0085 JVNGANESH"
>
> # 1. View basic summary
> print("Basic Summary of airquality dataset:")
[1] "Basic Summary of airquality dataset:"
> summary(airquality)
     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
>
> # 2. Data Transformation (Standardization and Log Transformation)
> airquality$Ozone_scaled <- as.numeric(scale(airquality$Ozone, center = TRUE, scale = TRUE))
> print("Standardized Ozone column:")
[1] "Standardized Ozone column:"
> print(head(airquality$Ozone_scaled, 10))
 [1] -0.03423409 -0.18580489 -0.91334473 -0.73145977          NA -0.42831817 -0.57988897 -0.70114561
 [9] -1.03460136          NA
> airquality$Log_Wind <- log(airquality$Wind + 1)
> print("Log transformed Wind column:")
[1] "Log transformed Wind column:"
> print(head(airquality$Log_Wind, 10))
 [1] 2.128232 2.197225 2.610070 2.525729 2.727853 2.766319 2.261763 2.694627 3.049273 2.261763
> # 3. Imputation of Missing Values
> airquality$Ozone[is.na(airquality$Ozone)] <- mean(airquality$Ozone, na.rm = TRUE)
> airquality$Solar.R[is.na(airquality$Solar.R)] <- median(airquality$Solar.R, na.rm = TRUE)
>
> airquality <- airquality %>%
+   group_by(Month) %>%
+   mutate(Ozone = ifelse(is.na(Ozone), median(Ozone, na.rm = TRUE), Ozone))
> # 4. Multiple Imputation Using MICE
> set.seed(123)  # Initialize random seed
> airquality_for_imputation <- airquality %>% select(-Ozone_scaled)
> airquality_imputed <- mice(airquality_for_imputation, method = 'pmm', m = 5)


 iter imp variable
  1   1
  1   2
  1   3
  1   4
  1   5
  2   1
  2   2
  2   3
  2   4

  2   5

  3   1

  3   2

  3   3

  3   4

  3   5

  4   1

  4   2

  4   3

  4   4

  4   5

  5   1

  5   2

  5   3

  5   4

  5   5

> completed_airquality <- complete(airquality_imputed)

> print("Completed airquality dataset after multiple imputation:")

[1] "Completed airquality dataset after multiple imputation:"

> print(completed_airquality)

      Ozone Solar.R Wind Temp Month Day  Log_Wind

1    41.00000     190  7.4   67     5   1 2.1282317

2    36.00000     118  8.0   72     5   2 2.1972246

3    12.00000     149 12.6   74     5   3 2.6100698
```

```
4   18.00000   313 11.5  62    5  4 2.5257286
5   42.12931   205 14.3  56    5  5 2.7278528
6   28.00000   205 14.9  66    5  6 2.7663191
7   23.00000   299  8.6  65    5  7 2.2617631
8   19.00000    99 13.8  59    5  8 2.6946272
9    8.00000    19 20.1  61    5  9 3.0492730
10  42.12931   194  8.6  69    5 10 2.2617631
11   7.00000   205  6.9  74    5 11 2.0668628
12  16.00000   256  9.7  69    5 12 2.3702437
13  11.00000   290  9.2  66    5 13 2.3223877
14  14.00000   274 10.9  68    5 14 2.4765384
15  18.00000    65 13.2  58    5 15 2.6532420
16  14.00000   334 11.5  64    5 16 2.5257286
17  34.00000   307 12.0  66    5 17 2.5649494
18   6.00000    78 18.4  57    5 18 2.9652731
19  30.00000   322 11.5  68    5 19 2.5257286
20  11.00000    44  9.7  62    5 20 2.3702437
21   1.00000     8  9.7  59    5 21 2.3702437
22  11.00000   320 16.6  73    5 22 2.8678989
23   4.00000    25  9.7  61    5 23 2.3702437
24  32.00000    92 12.0  61    5 24 2.5649494
25  42.12931    66 16.6  57    5 25 2.8678989
26  42.12931   266 14.9  58    5 26 2.7663191
27  42.12931   205  8.0  57    5 27 2.1972246
28  23.00000    13 12.0  67    5 28 2.5649494
```

```
29  45.00000 252 14.9 81 5 29 2.7663191
30 115.00000 223  5.7 79 5 30 1.9021075
31  37.00000 279  7.4 76 5 31 2.1282317
32  42.12931 286  8.6 78 6  1 2.2617631
33  42.12931 287  9.7 74 6  2 2.3702437
34  42.12931 242 16.1 67 6  3 2.8390785
35  42.12931 186  9.2 84 6  4 2.3223877
36  42.12931 220  8.6 85 6  5 2.2617631
37  42.12931 264 14.3 79 6  6 2.7278528
38  29.00000 127  9.7 82 6  7 2.3702437
39  42.12931 273  6.9 87 6  8 2.0668628
40  71.00000 291 13.8 90 6  9 2.6946272
41  39.00000 323 11.5 87 6 10 2.5257286
42  42.12931 259 10.9 93 6 11 2.4765384
43  42.12931 250  9.2 92 6 12 2.3223877
44  23.00000 148  8.0 82 6 13 2.1972246
45  42.12931 332 13.8 80 6 14 2.6946272
46  42.12931 322 11.5 79 6 15 2.5257286
47  21.00000 191 14.9 77 6 16 2.7663191
48  37.00000 284 20.7 72 6 17 3.0773123
49  20.00000  37  9.2 65 6 18 2.3223877
50  12.00000 120 11.5 73 6 19 2.5257286
51  13.00000 137 10.3 76 6 20 2.4248027
52  42.12931 150  6.3 77 6 21 1.9878743
53  42.12931  59  1.7 76 6 22 0.9932518
```

```
54  42.12931   91  4.6 76  6 23 1.7227666
55  42.12931  250  6.3 76  6 24 1.9878743
56  42.12931  135  8.0 75  6 25 2.1972246
57  42.12931  127  8.0 78  6 26 2.1972246
58  42.12931   47 10.3 73  6 27 2.4248027
59  42.12931   98 11.5 80  6 28 2.5257286
60  42.12931   31 14.9 77  6 29 2.7663191
61  42.12931  138  8.0 83  6 30 2.1972246
62 135.00000  269  4.1 84  7  1 1.6292405
63  49.00000  248  9.2 85  7  2 2.3223877
64  32.00000  236  9.2 81  7  3 2.3223877
65  42.12931  101 10.9 84  7  4 2.4765384
66  64.00000  175  4.6 83  7  5 1.7227666
67  40.00000  314 10.9 83  7  6 2.4765384
68  77.00000  276  5.1 88  7  7 1.8082888
69  97.00000  267  6.3 92  7  8 1.9878743
70  97.00000  272  5.7 92  7  9 1.9021075
71  85.00000  175  7.4 89  7 10 2.1282317
72  42.12931  139  8.6 82  7 11 2.2617631
73  10.00000  264 14.3 73  7 12 2.7278528
74  27.00000  175 14.9 81  7 13 2.7663191
75  42.12931  291 14.9 91  7 14 2.7663191
76   7.00000   48 14.3 80  7 15 2.7278528
77  48.00000  260  6.9 81  7 16 2.0668628
78  35.00000  274 10.3 82  7 17 2.4248027
```

```
79   61.00000   285  6.3  84   7 18 1.9878743
80   79.00000   187  5.1  87   7 19 1.8082888
81   63.00000   220 11.5  85   7 20 2.5257286
82   16.00000     7  6.9  74   7 21 2.0668628
83   42.12931   258  9.7  81   7 22 2.3702437
84   42.12931   295 11.5  82   7 23 2.5257286
85   80.00000   294  8.6  86   7 24 2.2617631
86  108.00000   223  8.0  85   7 25 2.1972246
87   20.00000    81  8.6  82   7 26 2.2617631
88   52.00000    82 12.0  86   7 27 2.5649494
89   82.00000   213  7.4  88   7 28 2.1282317
90   50.00000   275  7.4  86   7 29 2.1282317
91   64.00000   253  7.4  83   7 30 2.1282317
92   59.00000   254  9.2  81   7 31 2.3223877
93   39.00000    83  6.9  81   8  1 2.0668628
94    9.00000    24 13.8  81   8  2 2.6946272
95   16.00000    77  7.4  82   8  3 2.1282317
96   78.00000   205  6.9  86   8  4 2.0668628
97   35.00000   205  7.4  85   8  5 2.1282317
98   66.00000   205  4.6  87   8  6 1.7227666
99  122.00000   255  4.0  89   8  7 1.6094379
100  89.00000   229 10.3  90   8  8 2.4248027
101 110.00000   207  8.0  90   8  9 2.1972246
102  42.12931   222  8.6  92   8 10 2.2617631
103  42.12931   137 11.5  86   8 11 2.5257286
```

```
104  44.00000   192 11.5   86    8 12 2.5257286
105  28.00000   273 11.5   82    8 13 2.5257286
106  65.00000   157  9.7   80    8 14 2.3702437
107  42.12931    64 11.5   79    8 15 2.5257286
108  22.00000    71 10.3   77    8 16 2.4248027
109  59.00000    51  6.3   79    8 17 1.9878743
110  23.00000   115  7.4   76    8 18 2.1282317
111  31.00000   244 10.9   78    8 19 2.4765384
112  44.00000   190 10.3   78    8 20 2.4248027
113  21.00000   259 15.5   77    8 21 2.8033604
114   9.00000    36 14.3   72    8 22 2.7278528
115  42.12931   255 12.6   75    8 23 2.6100698
116  45.00000   212  9.7   79    8 24 2.3702437
117 168.00000   238  3.4   81    8 25 1.4816045
118  73.00000   215  8.0   86    8 26 2.1972246
119  42.12931   153  5.7   88    8 27 1.9021075
120  76.00000   203  9.7   97    8 28 2.3702437
121 118.00000   225  2.3   94    8 29 1.1939225
122  84.00000   237  6.3   96    8 30 1.9878743
123  85.00000   188  6.3   94    8 31 1.9878743
124  96.00000   167  6.9   91    9  1 2.0668628
125  78.00000   197  5.1   92    9  2 1.8082888
126  73.00000   183  2.8   93    9  3 1.3350011
127  91.00000   189  4.6   93    9  4 1.7227666
128  47.00000    95  7.4   87    9  5 2.1282317
```

```
129  32.00000     92 15.5   84     9   6 2.8033604
130  20.00000    252 10.9   80     9   7 2.4765384
131  23.00000    220 10.3   78     9   8 2.4248027
132  21.00000    230 10.9   75     9   9 2.4765384
133  24.00000    259  9.7   73     9  10 2.3702437
134  44.00000    236 14.9   81     9  11 2.7663191
135  21.00000    259 15.5   76     9  12 2.8033604
136  28.00000    238  6.3   77     9  13 1.9878743
137   9.00000     24 10.9   71     9  14 2.4765384
138  13.00000    112 11.5   71     9  15 2.5257286
139  46.00000    237  6.9   78     9  16 2.0668628
140  18.00000    224 13.8   67     9  17 2.6946272
141  13.00000     27 10.3   76     9  18 2.4248027
142  24.00000    238 10.3   68     9  19 2.4248027
 [ reached 'max' / getOption("max.print") -- omitted 11 rows ]
> # 5. Exploratory Data Analysis (EDA)
> summary(airquality)
     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 : 42.13   Median :205.0   Median : 9.700   Median :79.00   Median :7.000   Median :16.0
 Mean   : 42.13   Mean   :186.8   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
 Ozone_scaled       Log_Wind
 Min.   :-1.2468   Min.   :0.9933
 1st Qu.:-0.7315   1st Qu.:2.1282
 Median :-0.3222   Median :2.3702
 Mean   : 0.0000   Mean   :2.3376
 3rd Qu.: 0.6403   3rd Qu.:2.5257
 Max.   : 3.8157   Max.   :3.0773
 NA's   :37
>
> ggplot(airquality, aes(x = Wind, y = Ozone)) +
+   geom_point() +
+   geom_smooth(method = 'lm') +
+   labs(title = 'Scatter Plot of Wind vs Ozone')
`geom_smooth()` using formula = 'y ~ x'
> # 6. Data Cleaning and Outlier Detection
> Q1 <- quantile(airquality$Ozone, 0.25, na.rm = TRUE)
> Q3 <- quantile(airquality$Ozone, 0.75, na.rm = TRUE)
> IQR <- Q3 - Q1
> outliers <- which(airquality$Ozone < (Q1 - 1.5 * IQR) | airquality$Ozone > (Q3 + 1.5 * IQR))
> print("Detected Outliers:")
[1] "Detected Outliers:"
> print(airquality[outliers, ])
# A tibble: 15 × 8
# Groups:   Month [4]
  Ozone Solar.R  Wind  Temp Month   Day Ozone_scaled Log_Wind
  <dbl>   <dbl> <dbl> <int> <int> <int>       <dbl>    <dbl>
 1  115    223  5.7   79    5   30      2.21     1.90
 2  135    269  4.1   84    7    1      2.82    1.63
 3   97    267  6.3   92    7    8      1.66    1.99
 4   97    272  5.7   92    7    9      1.66    1.90
 5   85    175  7.4   89    7   10      1.30    2.13
 6  108    223  8     85    7   25      2.00    2.20
 7  122    255  4     89    8    7      2.42    1.61
 8   89    229 10.3   90    8    8      1.42    2.42
 9  110    207  8     90    8    9      2.06    2.20
10   168    238  3.4   81    8   25      3.82    1.48
11   118    225  2.3   94    8   29      2.30    1.19
12    84    237  6.3   96    8   30      1.27    1.99
13    85    188  6.3   94    8   31      1.30    1.99
14    96    167  6.9   91    9    1      1.63    2.07
15    91    189  4.6   93    9    4      1.48    1.72
>  7. Diagnose Missing Values using dlookr
Error: unexpected symbol in " 7. Diagnose"
> # 7. Diagnose Missing Values using dlookr
> airquality %>% diagnose() %>% print()
# A tibble: 40 × 8
  variables types   Month data_count missing_count missing_percent unique_count unique_rate
  <chr>     <chr>   <int>      <int>         <dbl>           <dbl>        <int>       <dbl>
 1 Ozone    numeric     5         31             0               0           22       0.710
 2 Ozone    numeric     6         30             0               0           10       0.333
 3 Ozone    numeric    7      31        0         0        25     0.806
 4 Ozone    numeric    8      31        0         0        25     0.806
 5 Ozone    numeric    9      30        0         0        22     0.733
 6 Solar.R  numeric    5      31        0         0        28     0.903
 7 Solar.R  numeric    6      30        0         0        28     0.933
 8 Solar.R  numeric    7      31        0         0        29     0.935
 9 Solar.R  numeric    8      31        0         0        28     0.903
10 Solar.R   numeric    9      30        0         0        27     0.9
# ℹ 30 more rows
# ℹ Use `print(n = ...)` to see more rows
>
> # Visualize missing data using naniar
> gg_miss_var(airquality)
> # 8. Regression-based Imputation
> ozone_model <- lm(Ozone ~ Wind + Temp, data = airquality)
> airquality$Ozone[is.na(airquality$Ozone)] <- predict(ozone_model, newdata = airquality[is.na(airquality
> print("After regression-based imputation of Ozone column:")
[1] "After regression-based imputation of Ozone column:"
> print(head(airquality$Ozone, 10))
 [1] 41.00000 36.00000 12.00000 18.00000 42.12931 28.00000 23.00000 19.00000  8.00000 42.12931
> # 9. Data Visualization after Imputation
> ggplot(airquality, aes(x = Ozone)) +
+   geom_histogram(bins = 30, fill = "blue", alpha = 0.7) +
+   labs(title = "Distribution of Ozone after Imputation")
>
> # Step 10: Additional Imputation Techniques
>
> # Imputation with Constant Value (Wind column)
> airquality$Wind <- impute(airquality$Wind, 5)
>
> # Visualize Missing Data Patterns using VIM
> aggr_plot <- aggr(airquality, col = c('navyblue', 'yellow'), numbers = TRUE, sortVars = TRUE, labels = nam
"Pattern"))


 Variables sorted by number of missings:
    Variable     Count
 Ozone_scaled 0.2418301
        Ozone 0.0000000
      Solar.R 0.0000000
         Wind 0.0000000
         Temp 0.0000000
        Month 0.0000000
          Day 0.0000000
     Log_Wind 0.0000000
>
> # Imputation for Entire Dataset using Median
> all_column_median <- apply(airquality, 2, median, na.rm = TRUE)
> for (i in colnames(airquality)) {
+   airquality[, i][is.na(airquality[, i])] <- all_column_median[i]
+ }
>
```

```
> # View the dataset after global median imputation

> print("Dataset after Global Median Imputation:")

[1] "Dataset after Global Median Imputation:"

> head(airquality)

# A tibble: 6 × 8

# Groups:   Month [1]

  Ozone Solar.R  Wind  Temp Month   Day Ozone_scaled Log_Wind

  <dbl>   <dbl> <dbl> <int> <int> <int>        <dbl>    <dbl>

1  41      190   7.4   67    5    1     -0.0342    2.13

2  36      118   8     72    5    2     -0.186     2.20

3  12      149  12.6   74    5    3     -0.913     2.61

4  18      313  11.5   62    5    4     -0.731     2.53

5  42.1    205  14.3   56    5    5     -0.322     2.73

6  28      205  14.9   66    5    6     -0.428     2.77
```