Yatish Kelkar TE IT 8001

DSBDA Assignment 6 : Air Quality Dataset

* Aim : Perform the following operations using R on the Air Quality data:

i) data cleaning
ii) data integration
iii) data transformation
iv) error correction
v) data model building

* Theory

Q1) What is data cleaning & data preparation?

Ans: i) Data Cleaning is the process of detecting & correcting & removing corrupt or inaccurate data.

ii) Data cleaning may be performed interactively with data wrangling tools or as batch processing through scripting.

iii) Data cleaning may also involve typographical errors or validating & correcting values. A common data cleaning practice is data enhancement where data is made more complete by adding information.

Q2) Explain the following in R :

Ans : a) na.omit() : Removes all incomplete cases of a data object (typically of a dataframe, matrix or vector).

Syntax : na.omit(data)

b) rbind() : Row binding joins multiple rows to form a single batch.

```
Syntax:  rbind(d1, d2)
Example:  d1 = data[1:5,]
          d2 = data[6:10,]
          d = rbind(d1, d2)
```

c) cbind() : Column binding is used to combine vectors, matrices and dataframes by colums.

```
Syntax :  cbind(d1, d2)
Example:  d1 = data[, 1:5]
          d2 = data[, 6:8]
          d = cbind(d1, d2)
```

* Conclusion : R Assignment for regression model on air quality dataset has been successfully implemented.

AirQuality.R

```
# Air Quality

# Yatish Kelkar TE IT 8001


data("airquality")

airQuality <- airquality


summary(airQuality)


# replace NA values with mean

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

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

summary(airQuality)


# data integration


subset1 <- airQuality[1:10, c(2,3)]

subset2 <- airQuality[1:10, c(4,5)]

cbind(subset1, subset2)


s1 <- airQuality[1:5, c(2,3,4,5)]

s2 <- airQuality[6:10, c(2,3,4,5)]

rbind(s1,s2)


# data transformation


copy <- airQuality
```

```
copy$Month <- month.abb[copy$Month]


# add a variable to check if solar value is dangerous


# airQuality$Dangerous <- airQuality$Solar.R > 110




# model building


plot(y~x)


#shuffle

set.seed(12345678)

airQuality <- airQuality[sample(nrow(airQuality)),]


splitPoint <- nrow(airQuality)*0.75

train <- airQuality[1:splitPoint,]

test <- airQuality[(splitPoint + 1):nrow(airQuality),]


train

test


model <- lm(Ozone~Solar.R, data = train)

model

abline(model, col="green", lwd = 5)


prediction <- predict(model, test)

prediction
```

Output

```
> # Air Quality
> # Yatish Kelkar TE IT 8001
>
> data("airquality")
> airQuality <- airquality
>
> summary(airQuality)
     Ozone           Solar.R           Wind             Temp      
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00  
 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00  
 Median : 31.50   Median :205.0   Median : 9.700   Median :79.00  
 Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88  
 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00  
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00  
 NA's   :37       NA's   :7                                       
     Month            Day      
 Min.   :5.000   Min.   : 1.0  
 1st Qu.:6.000   1st Qu.: 8.0  
 Median :7.000   Median :16.0  
 Mean   :6.993   Mean   :15.8  
 3rd Qu.:8.000   3rd Qu.:23.0  
 Max.   :9.000   Max.   :31.0  

>
> # replace NA values with mean
> airQuality$Ozone[is.na(airQuality$Ozone)] <- mean(airQuality$Ozone, na.r
m = TRUE)
> airQuality$Solar.R[is.na(airQuality$Solar.R)] <- mean(airQuality$Solar.R
, na.rm = TRUE)
>
> summary(airQuality)
     Ozone           Solar.R           Wind             Temp      
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00  
 1st Qu.: 21.00   1st Qu.:120.0   1st Qu.: 7.400   1st Qu.:72.00  
 Median : 42.13   Median :194.0   Median : 9.700   Median :79.00  
 Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88  
 3rd Qu.: 46.00   3rd Qu.:256.0   3rd Qu.:11.500   3rd Qu.:85.00  
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00  
     Month            Day      
 Min.   :5.000   Min.   : 1.0  
 1st Qu.:6.000   1st Qu.: 8.0  
 Median :7.000   Median :16.0  
 Mean   :6.993   Mean   :15.8  
 3rd Qu.:8.000   3rd Qu.:23.0  
 Max.   :9.000   Max.   :31.0  
>
> # data integration
>
> subset1 <- airQuality[1:10, c(2,3)]
> subset2 <- airQuality[1:10, c(4,5)]
> cbind(subset1, subset2)
    Solar.R Wind Temp Month
1  190.0000  7.4   67     5
2  118.0000  8.0   72     5
```

```
3  149.0000 12.6   74      5
4  313.0000 11.5   62      5
5  185.9315 14.3   56      5
6  185.9315 14.9   66      5
7  299.0000  8.6   65      5
8   99.0000 13.8   59      5
9   19.0000 20.1   61      5
10 194.0000  8.6   69      5
> 
> s1 <- airQuality[1:5, c(2,3,4,5)]
> s2 <- airQuality[6:10, c(2,3,4,5)]
> rbind(s1,s2)
    Solar.R Wind Temp Month
1  190.0000  7.4   67      5
2  118.0000  8.0   72      5
3  149.0000 12.6   74      5
4  313.0000 11.5   62      5
5  185.9315 14.3   56      5
6  185.9315 14.9   66      5
7  299.0000  8.6   65      5
8   99.0000 13.8   59      5
9   19.0000 20.1   61      5
10 194.0000  8.6   69      5
> 
> # data transformation
> 
> copy <- airQuality
> copy$Month <- month.abb[copy$Month]
> 
> # add a variable to check if solar value is dangerous
> 
> # airQuality$Dangerous <- airQuality$Solar.R > 110
> 
> 
> # model building
> 
> plot(y~x)
> 
> #shuffle
> set.seed(12345678)
> airQuality <- airQuality[sample(nrow(airQuality)),]
> 
> splitPoint <- nrow(airQuality)*0.75
> train <- airQuality[1:splitPoint,]
> test <- airQuality[(splitPoint + 1):nrow(airQuality),]
> 
> train
        Ozone  Solar.R Wind Temp Month Day
125  78.00000 197.0000  5.1   92     9   2
112  44.00000 190.0000 10.3   78     8  20
57   42.12931 127.0000  8.0   78     6  26
18    6.00000  78.0000 18.4   57     5  18
92   59.00000 254.0000  9.2   81     7  31
64   32.00000 236.0000  9.2   81     7   3
144  13.00000 238.0000 12.6   64     9  21
93   39.00000  83.0000  6.9   81     8   1
12   16.00000 256.0000  9.7   69     5  12
61   42.12931 138.0000  8.0   83     6  30
141  13.00000  27.0000 10.3   76     9  18
```

```
49   20.00000  37.0000  9.2   65      6  18
23    4.00000  25.0000  9.7   61      5  23
29   45.00000 252.0000 14.9   81      5  29
78   35.00000 274.0000 10.3   82      7  17
39   42.12931 273.0000  6.9   87      6   8
32   42.12931 286.0000  8.6   78      6   1
73   10.00000 264.0000 14.3   73      7  12
91   64.00000 253.0000  7.4   83      7  30
33   42.12931 287.0000  9.7   74      6   2
89   82.00000 213.0000  7.4   88      7  28
106  65.00000 157.0000  9.7   80      8  14
71   85.00000 175.0000  7.4   89      7  10
100  89.00000 229.0000 10.3   90      8   8
53   42.12931  59.0000  1.7   76      6  22
123  85.00000 188.0000  6.3   94      8  31
139  46.00000 237.0000  6.9   78      9  16
19   30.00000 322.0000 11.5   68      5  19
80   79.00000 187.0000  5.1   87      7  19
150  42.12931 145.0000 13.2   77      9  27
54   42.12931  91.0000  4.6   76      6  23
4    18.00000 313.0000 11.5   62      5   4
31   37.00000 279.0000  7.4   76      5  31
136  28.00000 238.0000  6.3   77      9  13
72   42.12931 139.0000  8.6   82      7  11
96   78.00000 185.9315  6.9   86      8   4
70   97.00000 272.0000  5.7   92      7   9
90   50.00000 275.0000  7.4   86      7  29
81   63.00000 220.0000 11.5   85      7  20
22   11.00000 320.0000 16.6   73      5  22
110  23.00000 115.0000  7.4   76      8  18
94    9.00000  24.0000 13.8   81      8   2
63   49.00000 248.0000  9.2   85      7   2
131  23.00000 220.0000 10.3   78      9   8
88   52.00000  82.0000 12.0   86      7  27
103  42.12931 137.0000 11.5   86      8  11
137   9.00000  24.0000 10.9   71      9  14
36   42.12931 220.0000  8.6   85      6   5
101 110.00000 207.0000  8.0   90      8   9
130  20.00000 252.0000 10.9   80      9   7
59   42.12931  98.0000 11.5   80      6  28
58   42.12931  47.0000 10.3   73      6  27
40   71.00000 291.0000 13.8   90      6   9
145  23.00000  14.0000  9.2   71      9  22
62  135.00000 269.0000  4.1   84      7   1
117 168.00000 238.0000  3.4   81      8  25
105  28.00000 273.0000 11.5   82      8  13
104  44.00000 192.0000 11.5   86      8  12
124  96.00000 167.0000  6.9   91      9   1
85   80.00000 294.0000  8.6   86      7  24
50   12.00000 120.0000 11.5   73      6  19
52   42.12931 150.0000  6.3   77      6  21
107  42.12931  64.0000 11.5   79      8  15
122  84.00000 237.0000  6.3   96      8  30
26   42.12931 266.0000 14.9   58      5  26
56   42.12931 135.0000  8.0   75      6  25
147   7.00000  49.0000 10.3   69      9  24
133  24.00000 259.0000  9.7   73      9  10
76    7.00000  48.0000 14.3   80      7  15
115  42.12931 255.0000 12.6   75      8  23
```

```
51   13.00000 137.0000 10.3   76     6  20
99  122.00000 255.0000  4.0   89     8   7
28   23.00000  13.0000 12.0   67     5  28
11    7.00000 185.9315  6.9   74     5  11
95   16.00000  77.0000  7.4   82     8   3
45   42.12931 332.0000 13.8   80     6  14
6    28.00000 185.9315 14.9   66     5   6
134  44.00000 236.0000 14.9   81     9  11
127  91.00000 189.0000  4.6   93     9   4
35   42.12931 186.0000  9.2   84     6   4
42   42.12931 259.0000 10.9   93     6  11
121 118.00000 225.0000  2.3   94     8  29
83   42.12931 258.0000  9.7   81     7  22
102  42.12931 222.0000  8.6   92     8  10
79   61.00000 285.0000  6.3   84     7  18
152  18.00000 131.0000  8.0   76     9  29
114   9.00000  36.0000 14.3   72     8  22
87   20.00000  81.0000  8.6   82     7  26
151  14.00000 191.0000 14.3   75     9  28
142  24.00000 238.0000 10.3   68     9  19
98   66.00000 185.9315  4.6   87     8   6
153  20.00000 223.0000 11.5   68     9  30
8    19.00000  99.0000 13.8   59     5   8
24   32.00000  92.0000 12.0   61     5  24
118  73.00000 215.0000  8.0   86     8  26
67   40.00000 314.0000 10.9   83     7   6
149  30.00000 193.0000  6.9   70     9  26
9     8.00000  19.0000 20.1   61     5   9
148  14.00000  20.0000 16.6   63     9  25
86  108.00000 223.0000  8.0   85     7  25
25   42.12931  66.0000 16.6   57     5  25
143  16.00000 201.0000  8.0   82     9  20
120  76.00000 203.0000  9.7   97     8  28
3    12.00000 149.0000 12.6   74     5   3
16   14.00000 334.0000 11.5   64     5  16
17   34.00000 307.0000 12.0   66     5  17
47   21.00000 191.0000 14.9   77     6  16
119  42.12931 153.0000  5.7   88     8  27
66   64.00000 175.0000  4.6   83     7   5
20   11.00000  44.0000  9.7   62     5  20
48   37.00000 284.0000 20.7   72     6  17
15   18.00000  65.0000 13.2   58     5  15
10   42.12931 194.0000  8.6   69     5  10
146  36.00000 139.0000 10.3   81     9  23
> test
        Ozone  Solar.R Wind Temp Month Day
7    23.00000 299.0000  8.6   65     5   7
132  21.00000 230.0000 10.9   75     9   9
82   16.00000   7.0000  6.9   74     7  21
77   48.00000 260.0000  6.9   81     7  16
30  115.00000 223.0000  5.7   79     5  30
111  31.00000 244.0000 10.9   78     8  19
108  22.00000  71.0000 10.3   77     8  16
68   77.00000 276.0000  5.1   88     7   7
27   42.12931 185.9315  8.0   57     5  27
135  21.00000 259.0000 15.5   76     9  12
46   42.12931 322.0000 11.5   79     6  15
60   42.12931  31.0000 14.9   77     6  29
13   11.00000 290.0000  9.2   66     5  13
```

```
21    1.00000   8.0000  9.7   59     5  21
41   39.00000 323.0000 11.5   87     6  10
140  18.00000 224.0000 13.8   67     9  17
43   42.12931 250.0000  9.2   92     6  12
34   42.12931 242.0000 16.1   67     6   3
97   35.00000 185.9315  7.4   85     8   5
75   42.12931 291.0000 14.9   91     7  14
1    41.00000 190.0000  7.4   67     5   1
38   29.00000 127.0000  9.7   82     6   7
44   23.00000 148.0000  8.0   82     6  13
5    42.12931 185.9315 14.3   56     5   5
84   42.12931 295.0000 11.5   82     7  23
69   97.00000 267.0000  6.3   92     7   8
37   42.12931 264.0000 14.3   79     6   6
55   42.12931 250.0000  6.3   76     6  24
116  45.00000 212.0000  9.7   79     8  24
109  59.00000  51.0000  6.3   79     8  17
128  47.00000  95.0000  7.4   87     9   5
138  13.00000 112.0000 11.5   71     9  15
126  73.00000 183.0000  2.8   93     9   3
113  21.00000 259.0000 15.5   77     8  21
2    36.00000 118.0000  8.0   72     5   2
129  32.00000  92.0000 15.5   84     9   6
14   14.00000 274.0000 10.9   68     5  14
74   27.00000 175.0000 14.9   81     7  13
>
> model <- lm(Ozone~Solar.R, data = train)
> model

Call:
lm(formula = Ozone ~ Solar.R, data = train)

Coefficients:
(Intercept)      Solar.R
    21.6852       0.1188

> abline(model, col="green", lwd = 5)
>
> prediction <- predict(model, test)
> prediction
       7      132       82       77       30      111      108       68
57.22128 49.02063 22.51710 52.58613 48.18868 50.68453 30.12350 54.48773
      27      135       46       60       13       21       41      140
43.78310 52.46728 59.95483 25.36950 56.15163 22.63595 60.07368 48.30753
      43       34       97       75        1       38       44        5
51.39763 50.44683 43.78310 56.27048 44.26664 36.77909 39.27494 43.78310
      84       69       37       55      116      109      128      138
56.74588 53.41808 53.06153 51.39763 46.88134 27.74650 32.97589 34.99634
     126      113        2      129       14       74
43.43469 52.46728 35.70944 32.61934 54.25003 42.48389
```