Homework-1-R-Exercises

VISHAL BHASHYAAM

Table of contents

0.1 Getting to know the Data with R

0.1.1 Goal:

Practice basic R commands/methods for descriptive data analysis.

0.2 Installing required packages

```r
# run install.packages if package not downloaded
if(!require("pacman"))
  install.packages("pacman")
```

```
Loading required package: pacman
```

```r
library(pacman)

p_load(dlookr,
  DMwR2,
  GGally,
  Hmisc,
  palmerpenguins,
  tidyverse
)
```

Loading data

The |> is the Base R pipe as opposed to the magrittr pipe %>%. The |> pipe can be utilized for most functions in R, while the %>% pipe is more restricted towards the tidyverse

```r
data(algae, package ="DMwR2")
algae |> glimpse()
```

```
Rows: 200
Columns: 18
$ season <fct> winter, spring, autumn, spring, autumn, winter, summer, autumn,~
$ size   <fct> small, small, small, small, small, small, small, small, small, ~
$ speed  <fct> medium, medium, medium, medium, medium, high, high, high, mediu~
$ mxPH   <dbl> 8.00, 8.35, 8.10, 8.07, 8.06, 8.25, 8.15, 8.05, 8.70, 7.93, 7.7~
$ mnO2   <dbl> 9.8, 8.0, 11.4, 4.8, 9.0, 13.1, 10.3, 10.6, 3.4, 9.9, 10.2, 11.~
$ Cl     <dbl> 60.800, 57.750, 40.020, 77.364, 55.350, 65.750, 73.250, 59.067,~
$ NO3    <dbl> 6.238, 1.288, 5.330, 2.302, 10.416, 9.248, 1.535, 4.990, 0.886,~
$ NH4    <dbl> 578.000, 370.000, 346.667, 98.182, 233.700, 430.000, 110.000, 2~
$ oPO4   <dbl> 105.000, 428.750, 125.667, 61.182, 58.222, 18.250, 61.250, 44.6~
$ PO4    <dbl> 170.000, 558.750, 187.057, 138.700, 97.580, 56.667, 111.750, 77~
$ Chla   <dbl> 50.000, 1.300, 15.600, 1.400, 10.500, 28.400, 3.200, 6.900, 5.5~
$ a1     <dbl> 0.0, 1.4, 3.3, 3.1, 9.2, 15.1, 2.4, 18.2, 25.4, 17.0, 16.6, 32.~
$ a2     <dbl> 0.0, 7.6, 53.6, 41.0, 2.9, 14.6, 1.2, 1.6, 5.4, 0.0, 0.0, 0.0, ~
$ a3     <dbl> 0.0, 4.8, 1.9, 18.9, 7.5, 1.4, 3.2, 0.0, 2.5, 0.0, 0.0, 0.0, 2.~
$ a4     <dbl> 0.0, 1.9, 0.0, 0.0, 0.0, 0.0, 3.9, 0.0, 0.0, 2.9, 0.0, 0.0, 0.0~
$ a5     <dbl> 34.2, 6.7, 0.0, 1.4, 7.5, 22.5, 5.8, 5.5, 0.0, 0.0, 1.2, 0.0, 1~
$ a6     <dbl> 8.3, 0.0, 0.0, 0.0, 4.1, 12.6, 6.8, 8.7, 0.0, 0.0, 0.0, 0.0, 0.~
$ a7     <dbl> 0.0, 2.1, 9.7, 1.4, 1.0, 2.9, 0.0, 0.0, 0.0, 1.7, 6.0, 1.5, 2.1~
```

1 Central tendency: mean, median, mode

1.1 Mean

```
algae$a1 |>
  mean()
```

```
[1] 16.9235
```

1.2 Median

```
algae$a1 |>
  median()
```

```
[1] 6.95
```

1.3 Mode

Base R doesn't have a function for mode,

Creating a R function for mode, (works for unimodal, bimodal, multimodal data)

```r
Mode <- function(x, na.rm=FALSE){
  if (na.rm) x<-x[!is.na(x)]
  ux <- unique(x)
  return(ux[which.max(tabulate(match(x,ux)))])
}

algae$mnO2 |> Mode()
```

```
[1] 9.8
```

1.4 DMwR centralValue() function:

returns the median for numerical variable, or the mode for nominal variables.

```r
# Numerical variable
algae$a1 |> centralValue()
```

```
[1] 6.95
```

```r
# Nominal variable
algae$speed |> centralValue()
```

```
[1] "high"
```

1.5 Statistics of spread (variation)

1.6 Variance

```r
algae$a3 |> var()
```

```
[1] 48.28217
```

1.7 Standard deviation

```
algae$a3 |> sd()
```

```
[1] 6.948537
```

1.8 Range

Note that this gives you both maximum and minimum values.

```
algae$a4 |> range()
```

```
[1]  0.0 44.6
```

1.9 Maximum value

```
algae$a1 |> max ()
```

```
[1] 89.8
```

1.10 Minimum value

```
algae$a1 |> min()
```

```
[1] 0
```

1.11 Interquartile range

3rd quartile (75%) - 1st quartile (25%)

```
algae$a1 |> IQR()
```

```
[1] 23.3
```

1.12 Quantiles

```
algae$a1 |> quantile()
```

```
   0%   25%   50%   75%  100% 
 0.00  1.50  6.95 24.80 89.80 
```

Specifying particular quantiles:

```
algae$a1 |> quantile(probs = c(0.2,0.8))
```

```
  20%   80% 
 1.20 32.18 
```

1.13 Missing values

```
library(purrr)
#compute the total number of NA values in the given dataset

na_value <- algae %>%
  purrr::map_dbl(~sum(is.na(.))) %>%
  sum()

cat("The dataset contains ", na_value, "NA values. \n" )
```

```
The dataset contains  33 NA values. 
```

```
# Compute the number of incomplete rows in the dataset


  incomplete_rows <- algae %>%
  summarise_all(~!complete.cases(.)) %>%
  nrow()
```

```
Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in
dplyr 1.1.0.
i Please use `reframe()` instead.
```

```
i When switching from `summarise()` to `reframe()`, remember that `reframe()`
  always returns an ungrouped data frame and adjust accordingly.
i The deprecated feature was likely used in the dplyr package.
  Please report the issue at <https://github.com/tidyverse/dplyr/issues>.
```

```
cat("The dataset contains ", incomplete_rows, "(out of ", nrow(algae),") incomplete rows.
```

```
The dataset contains  200 (out of  200 ) incomplete rows.
```

2 Summaries of a dataset

2.1 Baser R's summary()

```
algae |> summary()
```

```
    season        size        speed         mxPH            mnO2       
 autumn:40   large :45   high  :84   Min.   :5.600   Min.   : 1.500  
 spring:53   medium:84   low   :33   1st Qu.:7.700   1st Qu.: 7.725  
 summer:45   small :71   medium:83   Median :8.060   Median : 9.800  
 winter:62                           Mean   :8.012   Mean   : 9.118  
                                     3rd Qu.:8.400   3rd Qu.:10.800  
                                     Max.   :9.700   Max.   :13.400  
                                     NA's   :1       NA's   :2       
       Cl               NO3              NH4                oPO4       
 Min.   :  0.222   Min.   : 0.050   Min.   :    5.00   Min.   :  1.00  
 1st Qu.: 10.981   1st Qu.: 1.296   1st Qu.:   38.33   1st Qu.: 15.70  
 Median : 32.730   Median : 2.675   Median :  103.17   Median : 40.15  
 Mean   : 43.636   Mean   : 3.282   Mean   :  501.30   Mean   : 73.59  
 3rd Qu.: 57.824   3rd Qu.: 4.446   3rd Qu.:  226.95   3rd Qu.: 99.33  
 Max.   :391.500   Max.   :45.650   Max.   :24064.00   Max.   :564.60  
 NA's   :10        NA's   :2        NA's   :2          NA's   :2       
      PO4              Chla               a1              a2        
 Min.   :  1.00   Min.   :  0.200   Min.   : 0.00   Min.   : 0.000  
 1st Qu.: 41.38   1st Qu.:  2.000   1st Qu.: 1.50   1st Qu.: 0.000  
 Median :103.29   Median :  5.475   Median : 6.95   Median : 3.000  
 Mean   :137.88   Mean   : 13.971   Mean   :16.92   Mean   : 7.458  
 3rd Qu.:213.75   3rd Qu.: 18.308   3rd Qu.:24.80   3rd Qu.:11.375  
 Max.   :771.60   Max.   :110.456   Max.   :89.80   Max.   :72.600  
 NA's   :2        NA's   :12
       a3               a4               a5               a6
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
 Median : 1.550   Median : 0.000   Median : 1.900   Median : 0.000
 Mean   : 4.309   Mean   : 1.992   Mean   : 5.064   Mean   : 5.964
 3rd Qu.: 4.925   3rd Qu.: 2.400   3rd Qu.: 7.500   3rd Qu.: 6.925
 Max.   :42.800   Max.   :44.600   Max.   :44.400   Max.   :77.600

       a7
 Min.   : 0.000
 1st Qu.: 0.000
 Median : 1.000
 Mean   : 2.495
 3rd Qu.: 2.400
 Max.   :31.600
```

2.2 Hmisc's describe()

```r
data("penguins")
penguins |> Hmisc::describe()
```

```
penguins

 8  Variables      344  Observations
--------------------------------------------------------------------------------
species
       n  missing distinct
     344        0        3

Value         Adelie Chinstrap    Gentoo
Frequency        152        68       124
Proportion     0.442     0.198     0.360
--------------------------------------------------------------------------------
island
       n  missing distinct
     344        0        3

Value         Biscoe     Dream Torgersen
Frequency        168       124        52
```

```
Proportion     0.488     0.360     0.151
--------------------------------------------------------------------------------
bill_length_mm
       n  missing distinct     Info     Mean      Gmd      .05      .10
     342        2      164        1    43.92    6.274    35.70    36.60
     .25      .50      .75      .90      .95
   39.23    44.45    48.50    50.80    51.99

lowest : 32.1 33.1 33.5 34   34.1, highest: 55.1 55.8 55.9 58   59.6
--------------------------------------------------------------------------------
bill_depth_mm
       n  missing distinct     Info     Mean      Gmd      .05      .10
     342        2       80        1    17.15    2.267     13.9     14.3
     .25      .50      .75      .90      .95
    15.6     17.3     18.7     19.5     20.0

lowest : 13.1 13.2 13.3 13.4 13.5, highest: 20.7 20.8 21.1 21.2 21.5
--------------------------------------------------------------------------------
flipper_length_mm
       n  missing distinct     Info     Mean      Gmd      .05      .10
     342        2       55    0.999    200.9    16.03    181.0    185.0
     .25      .50      .75      .90      .95
   190.0    197.0    213.0    220.9    225.0

lowest : 172 174 176 178 179, highest: 226 228 229 230 231
--------------------------------------------------------------------------------
body_mass_g
       n  missing distinct     Info     Mean      Gmd      .05      .10
     342        2       94        1     4202    911.8     3150     3300
     .25      .50      .75      .90      .95
    3550     4050     4750     5400     5650

lowest : 2700 2850 2900 2925 2975, highest: 5850 5950 6000 6050 6300
--------------------------------------------------------------------------------
sex
       n  missing distinct
     333       11        2

Value      female   male
Frequency     165    168
Proportion  0.495  0.505
--------------------------------------------------------------------------------
year
```

```
       n  missing distinct     Info     Mean      Gmd
     344        0        3    0.888     2008   0.8919

Value        2007  2008  2009
Frequency     110   114   120
Proportion  0.320 0.331 0.349

For the frequency table, variable is rounded to the nearest 0.02
--------------------------------------------------------------------------------
```

GMD is the mean absolute difference between any pairs of observations. A robust dispersion measure, especially for non- normally distributed data.

2.3 dlookr's describe()

```
penguins |> dlookr::describe()
```

```
# A tibble: 5 x 26
  described_variables     n    na   mean      sd se_mean     IQR skewness
  <chr>               <int> <int>  <dbl>   <dbl>   <dbl>   <dbl>    <dbl>
1 bill_length_mm        342     2   43.9   5.46   0.295     9.27   0.0531
2 bill_depth_mm         342     2   17.2   1.97   0.107     3.1   -0.143
3 flipper_length_mm     342     2  201.   14.1    0.760    23       0.346
4 body_mass_g           342     2 4202.  802.    43.4    1200       0.470
5 year                  344     0 2008.    0.818  0.0441    2     -0.0537
# i 18 more variables: kurtosis <dbl>, p00 <dbl>, p01 <dbl>, p05 <dbl>,
#   p10 <dbl>, p20 <dbl>, p25 <dbl>, p30 <dbl>, p40 <dbl>, p50 <dbl>,
#   p60 <dbl>, p70 <dbl>, p75 <dbl>, p80 <dbl>, p90 <dbl>, p95 <dbl>,
#   p99 <dbl>, p100 <dbl>
```

2.4 Summaries on a subset of data

`dplyr` 's `summarise()` and `summarise_all()` or use them with `select()` and `group_by()` to create summaries on subset of data. And,

`summarise()`= `summarize()`

```
algae |>
  summarize(avgNO3 = mean(NO3,na.rm=TRUE),
            medA1 = median(a1))
```

```
# A tibble: 1 x 2
  avgNO3 medA1
   <dbl> <dbl>
1   3.28  6.95
```

summarize_all() can be used to apply any function that produces a scalar value to any column of a data

```
algae |>
  select(mxPH:Cl) |>
  summarize_all(list(mean,median), na.rm=TRUE)
```

```
# A tibble: 1 x 6
  mxPH_fn1 mnO2_fn1 Cl_fn1 mxPH_fn2 mnO2_fn2 Cl_fn2
     <dbl>    <dbl>  <dbl>    <dbl>    <dbl>  <dbl>
1     8.01     9.12   43.6     8.06      9.8   32.7
```

```
algae |>
  select(a1:a7) |>
  summarize_all(funs(var))
```

```
Warning: `funs()` was deprecated in dplyr 0.8.0.
i Please use a list of either functions or lambdas:

# Simple named list: list(mean = mean, median = median)

# Auto named with `tibble::lst()`: tibble::lst(mean, median)

# Using lambdas list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
```

```
# A tibble: 1 x 7
     a1    a2    a3    a4    a5    a6    a7
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1  456.  122.  48.3  19.5  56.1  136.  26.6
```

```
algae |>
  select (a1:a7) |>
  summarise_all(c("min","max"))
```

```
# A tibble: 1 x 14
  a1_min a2_min a3_min a4_min a5_min a6_min a7_min a1_max a2_max a3_max a4_max
   <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
1      0      0      0      0      0      0      0   89.8   72.6   42.8   44.6
# i 3 more variables: a5_max <dbl>, a6_max <dbl>, a7_max <dbl>
```

2.5 Use summarize() and group_by()

```
algae |>
  group_by(season, size) |>
  summarize(nObs = n(), mA7=median(a7))
```

```
`summarise()` has grouped output by 'season'. You can override using the
`.groups` argument.
```

```
# A tibble: 12 x 4
# Groups:   season [4]
   season size    nObs   mA7
   <fct>  <fct>  <int> <dbl>
 1 autumn large     11  0
 2 autumn medium    16  1.05
 3 autumn small     13  0
 4 spring large     12  1.95
 5 spring medium    21  1
 6 spring small     20  0
 7 summer large     10  0
 8 summer medium    21  1
 9 summer small     14  1.45
10 winter large     12  0
11 winter medium    26  1.4
12 winter small     24  0
```

```
penguins |>
  group_by(species) |>
  summarize(var = var(bill_length_mm, na.rm = TRUE))
```

```
# A tibble: 3 x 2
  species     var
  <fct>     <dbl>
```

```
1 Adelie     7.09
2 Chinstrap 11.2
3 Gentoo     9.50
```

2.6 Aggregating data

Useful for summary function that don't return scalar values

```
penguins |>
  group_by(species) |>
  reframe(var = quantile(bill_length_mm, na.rm = TRUE))
```

```
# A tibble: 15 x 2
   species     var
   <fct>     <dbl>
 1 Adelie     32.1
 2 Adelie     36.8
 3 Adelie     38.8
 4 Adelie     40.8
 5 Adelie     46
 6 Chinstrap  40.9
 7 Chinstrap  46.3
 8 Chinstrap  49.6
 9 Chinstrap  51.1
10 Chinstrap  58
11 Gentoo     40.9
12 Gentoo     45.3
13 Gentoo     47.3
14 Gentoo     49.6
15 Gentoo     59.6
```

`reframe()` expectsd a scalar result returned by the function, but quantile returns a vector.

Aggregating data with summarize was depreciated in `dplyr 1.1.0` , `reframe()` should be used instead.

```
penguins |>
  group_by(species) |>
  dlookr::describe(bill_length_mm)
```

```
# A tibble: 3 x 27
  described_variables species       n    na  mean    sd se_mean   IQR skewness
  <chr>               <fct>     <int> <int> <dbl> <dbl>   <dbl> <dbl>    <dbl>
1 bill_length_mm      Adelie      151     1  38.8  2.66   0.217  4       0.162
2 bill_length_mm      Chinstrap    68     0  48.8  3.34   0.405  4.73 -0.0906
3 bill_length_mm      Gentoo      123     1  47.5  3.08   0.278  4.25   0.651
# i 18 more variables: kurtosis <dbl>, p00 <dbl>, p01 <dbl>, p05 <dbl>,
#   p10 <dbl>, p20 <dbl>, p25 <dbl>, p30 <dbl>, p40 <dbl>, p50 <dbl>,
#   p60 <dbl>, p70 <dbl>, p75 <dbl>, p80 <dbl>, p90 <dbl>, p95 <dbl>,
#   p99 <dbl>, p100 <dbl>
```

2.6.1 List data types of the attributes in tidy dataset

```
str(algae)# display data types
```

```
tibble [200 x 18] (S3: tbl_df/tbl/data.frame)
 $ season: Factor w/ 4 levels "autumn","spring",..: 4 2 1 2 1 4 3 1 4 4 ...
 $ size  : Factor w/ 3 levels "large","medium",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ speed : Factor w/ 3 levels "high","low","medium": 3 3 3 3 3 1 1 1 3 1 ...
 $ mxPH  : num [1:200] 8 8.35 8.1 8.07 8.06 8.25 8.15 8.05 8.7 7.93 ...
 $ mnO2  : num [1:200] 9.8 8 11.4 4.8 9 13.1 10.3 10.6 3.4 9.9 ...
 $ Cl    : num [1:200] 60.8 57.8 40 77.4 55.4 ...
 $ NO3   : num [1:200] 6.24 1.29 5.33 2.3 10.42 ...
 $ NH4   : num [1:200] 578 370 346.7 98.2 233.7 ...
 $ oPO4  : num [1:200] 105 428.8 125.7 61.2 58.2 ...
 $ PO4   : num [1:200] 170 558.8 187.1 138.7 97.6 ...
 $ Chla  : num [1:200] 50 1.3 15.6 1.4 10.5 ...
 $ a1    : num [1:200] 0 1.4 3.3 3.1 9.2 15.1 2.4 18.2 25.4 17 ...
 $ a2    : num [1:200] 0 7.6 53.6 41 2.9 14.6 1.2 1.6 5.4 0 ...
 $ a3    : num [1:200] 0 4.8 1.9 18.9 7.5 1.4 3.2 0 2.5 0 ...
 $ a4    : num [1:200] 0 1.9 0 0 0 0 3.9 0 0 2.9 ...
 $ a5    : num [1:200] 34.2 6.7 0 1.4 7.5 22.5 5.8 5.5 0 0 ...
 $ a6    : num [1:200] 8.3 0 0 0 4.1 12.6 6.8 8.7 0 0 ...
 $ a7    : num [1:200] 0 2.1 9.7 1.4 1 2.9 0 0 0 1.7 ...
```

```
Hmisc::describe(algae) # description of the values
```

```
algae
```

```
 18  Variables      200  Observations
--------------------------------------------------------------------------------
season
       n  missing distinct
     200        0        4

Value      autumn spring summer winter
Frequency      40     53     45     62
Proportion  0.200  0.265  0.225  0.310
--------------------------------------------------------------------------------
size
       n  missing distinct
     200        0        3

Value       large medium  small
Frequency      45     84     71
Proportion  0.225  0.420  0.355
--------------------------------------------------------------------------------
speed
       n  missing distinct
     200        0        3

Value        high    low medium
Frequency      84     33     83
Proportion  0.420  0.165  0.415
--------------------------------------------------------------------------------
mxPH
       n  missing distinct     Info     Mean      Gmd      .05      .10
     199        1       72    0.998    8.012   0.6471    7.081    7.340
     .25      .50      .75      .90      .95
   7.700    8.060    8.400    8.700    8.873

lowest : 5.6  5.7  6.4  6.5  6.6 , highest: 9    9.06 9.1  9.5  9.7
--------------------------------------------------------------------------------
mnO2
       n  missing distinct     Info     Mean      Gmd      .05      .10
     198        2       88        1    9.118    2.629    4.485    5.770
     .25      .50      .75      .90      .95
   7.725    9.800   10.800   11.700   11.815

lowest : 1.5  1.8  3.2  3.3  3.4 , highest: 12.5 12.6 12.9 13.1 13.4
--------------------------------------------------------------------------------
Cl
```

```
       n  missing distinct     Info     Mean      Gmd      .05      .10
     190       10      178        1    43.64    43.78    3.061    4.970
     .25      .50      .75      .90      .95
  10.981   32.730   57.823   88.600  130.087

lowest : 0.222   0.8     1.17    1.45    1.549
highest: 173.75  187.183 194.75  208.364 391.5
--------------------------------------------------------------------------------
NO3
       n  missing distinct     Info     Mean      Gmd      .05      .10
     198        2      192        1    3.282    2.884   0.4023   0.6912
     .25      .50      .75      .90      .95
  1.2960   2.6750   4.4463   6.1916   7.9369

lowest : 0.05   0.102  0.13   0.23   0.267 , highest: 9.248  9.715  9.773  10.416 45.65
--------------------------------------------------------------------------------
NH4
       n  missing distinct     Info     Mean      Gmd      .05      .10
     198        2      179        1    501.3    816.2    10.00    15.00
     .25      .50      .75      .90      .95
   38.33   103.17   226.95   805.33  1922.87

lowest : 5       5.8     8       10      10.5
highest: 4073.33 5738.33 6400    8777.6  24064
--------------------------------------------------------------------------------
oPO4
       n  missing distinct     Info     Mean      Gmd      .05      .10
     198        2      173        1    73.59    85.46     2.00     3.94
     .25      .50      .75      .90      .95
   15.70    40.15    99.33   193.21   248.34

lowest : 1       1.25    1.333   1.625   1.8
highest: 346.167 412.333 428.75  467.5   564.6
--------------------------------------------------------------------------------
PO4
       n  missing distinct     Info     Mean      Gmd      .05      .10
     198        2      189        1    137.9    133.9    6.455   11.350
     .25      .50      .75      .90      .95
  41.375  103.285  213.750  286.100  345.650

lowest : 1       2.5     3       4       6
highest: 558.75  586     607.167 624.733 771.6
--------------------------------------------------------------------------------
```

```
Chla
       n  missing distinct     Info     Mean      Gmd      .05      .10
     188       12      131        1    13.97    17.93    0.500    0.800
     .25      .50      .75      .90      .95
   2.000    5.475   18.308   31.817   61.733

lowest : 0.2     0.3     0.4     0.5     0.6
highest: 88.255  92.667  93.683  98.817  110.456
--------------------------------------------------------------------------------
a1
       n  missing distinct     Info     Mean      Gmd      .05      .10
     200        0      121    0.994    16.92    21.52     0.00     0.00
     .25      .50      .75      .90      .95
    1.50     6.95    24.80    50.72    64.33

lowest : 0    1.1  1.2  1.4  1.5 , highest: 75.8 81.9 82.7 86.6 89.8
--------------------------------------------------------------------------------
a2
       n  missing distinct     Info     Mean      Gmd      .05      .10
     200        0       89    0.951    7.458    10.19     0.00     0.00
     .25      .50      .75      .90      .95
    0.00     3.00    11.38    21.50    28.38

lowest : 0    1    1.2  1.4  1.5 , highest: 40.7 40.9 41   53.6 72.6
--------------------------------------------------------------------------------
a3
       n  missing distinct     Info     Mean      Gmd      .05      .10
     200        0       79    0.949    4.309    6.131    0.000    0.000
     .25      .50      .75      .90      .95
   0.000    1.550    4.925   13.510   20.275

lowest : 0    1    1.1  1.2  1.4 , highest: 24.8 25.3 25.9 35.1 42.8
--------------------------------------------------------------------------------
a4
       n  missing distinct     Info     Mean      Gmd      .05      .10
     200        0       50    0.838    1.992    3.032    0.000    0.000
     .25      .50      .75      .90      .95
   0.000    0.000    2.400    5.000    7.605

lowest : 0    1    1.1  1.2  1.3 , highest: 11.5 12.7 13.4 28.8 44.6
--------------------------------------------------------------------------------
a5
       n  missing distinct     Info     Mean      Gmd      .05      .10
     200        0       81    0.938    5.064    6.923     0.00     0.00
     .25      .50      .75      .90      .95
    0.00     1.90     7.50    14.91    20.04

lowest : 0    1    1.1  1.2  1.4 , highest: 28.8 34.2 34.3 35.6 44.4
--------------------------------------------------------------------------------
a6
       n  missing distinct     Info     Mean      Gmd      .05      .10
     200        0       76    0.847    5.964    9.323    0.000    0.000
     .25      .50      .75      .90      .95
   0.000    0.000    6.925   17.110   31.815

lowest : 0    1    1.2  1.4  1.5 , highest: 42.7 49.4 52.5 64.6 77.6
--------------------------------------------------------------------------------
a7
       n  missing distinct     Info     Mean      Gmd      .05      .10
     200        0       51    0.882    2.496    3.817     0.00     0.00
     .25      .50      .75      .90      .95
    0.00     1.00     2.40     6.10    10.88

lowest : 0    1    1.1  1.2  1.4 , highest: 22.1 25.6 30.1 31.2 31.6
--------------------------------------------------------------------------------
```

2.6.2 Check skewness in data distribution in attributes

Use "`skewness()`" from `e1071` package to find the skewness in data distribution.

```r
if(!require("e1071"))
  install.packages("e1071")
```

```
Loading required package: e1071

Attaching package: 'e1071'

The following object is masked from 'package:Hmisc':

    impute

The following objects are masked from 'package:dlookr':

    kurtosis, skewness
```

```
library(e1071)

skewValue<- skewness(algae$a2)
cat("Skewness value is, ", skewValue)
```

```
Skewness value is,  2.395171
```

2.7 Correlation

```
# Calculate correlations for numeric columns in the dataset
correlation_value <- cor(algae$a1, algae$a2)


cat("correlation between a1 and a2 : ",correlation_value)
```

```
correlation between a1 and a2 :  -0.2937678
```

2.8 Examine number of missing values in dataset

```
cat("missing values in algae dataset is : ", sum(is.na(algae)))
```

```
missing values in algae dataset is :  33
```

2.9 Ways to overcome missing values:

- Either the NA values can be omitted using "`na.omit()`"

```
algae_data<- na.omit(algae)
cat("missing values in algae dataset is : ", sum(is.na(algae_data)))
```

```
missing values in algae dataset is :  0
```

- Else we can take the average of the particular column to fill the NA values using `mean()`

```
is.na(algae$Cl)
```

```
  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [49] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
 [61]  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[157] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[169] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[181] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[193] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
```

```
algae_1 <- algae
algae_1$Cl[is.na(algae_1$Cl)]<-mean(algae_1$Cl,na.rm=TRUE)

algae_1$Cl
```

```
  [1]  60.80000  57.75000  40.02000  77.36400  55.35000  65.75000  73.25000
  [8]  59.06700  21.95000   8.00000   8.00000   8.69000   5.00000   6.30000
 [15]   3.00000   4.70000   7.00000   7.00000   7.00000  64.00000  88.00000
 [22]   0.80000  32.92000  11.86700  10.97500  12.53600  10.50000   9.00000
 [29]  16.00000   9.00000  13.00000  26.00000  20.08300  34.50000  29.20000
 [36]  30.52300   1.17000   1.45000  20.62500  22.28600  77.00000  54.19000
 [43]  50.00000  54.14300  69.75000  87.00000  66.30000   9.00000  15.00000
 [50]  17.75000  32.30000  27.23300   6.16700   5.27300  43.63628  43.63628
 [57]  43.63628  43.63628  43.63628  43.63628  43.63628  43.63628   4.08300
 [64]   4.57500   4.32600   2.93300   3.27500   3.13600  32.40000  29.77500
 [71]  32.54000  38.12500  34.03700 136.00000 129.37500  35.75000  29.50000
 [78]  27.40000  26.76000  11.00000  11.00000  10.40000  13.50000  12.14600
 [85]  31.00000  53.00000  36.24800  48.66700  53.10200 125.60000 173.75000
 [92]  94.40500  53.33300  70.00000  63.51000  56.71700  61.05000  57.75000
 [99] 101.87500  85.98200  63.62500  82.11100  65.33300  58.33100  49.62500
[106]  47.77800  47.22900  41.50000  40.16700  32.05600   5.88900   7.25000
[113]   7.83800  53.42500  57.84800   0.22200   1.54900   5.83000  74.66700
[120] 131.39999  45.27300  42.63600  48.42900  11.81800  10.55600  12.00000
```

```
[127]  31.09100  28.33300  30.12500  10.93600  10.07800  11.08800 194.75000
[134] 391.50000 130.67000  39.00000  35.66000  37.60000  39.00000  49.90000
[141]  51.11300   8.30000  10.20700  79.07700  81.33300  64.09300  41.25000
[148]  40.22600  46.16700  47.00000  41.16300  53.00000  44.20500 127.83300
[155] 100.83000  94.00000  69.00000  50.00000  19.22000  26.00000  43.63628
[162]  44.00000  43.00000  43.09000  16.00000  22.35000  82.85700  63.29200
[169]  43.97000  38.90200  95.36700 151.83299 104.81800  71.44400 208.36400
[176] 187.18300   4.54500   3.50000   5.32600   2.11100   2.20000   2.75000
[183]   3.86000   9.05500   7.61300  39.10900  22.45500  23.25000  22.32000
[190]  12.77800  15.54100  12.18200   7.33300  23.82500  12.44400  17.37500
[197]  14.32000 139.98900  43.63628  82.85200
```