R Programming and Bio-conductor Assignment Report

Mutesasira Edward

Date: 2023-10-30

Section A: Data Exploration 1

metaFile.csv file was read into R session and the following were the explorations.
The number of records and columns.

```
[1] 50  3
```

The number of samples which were obtained for each level of treatment

```
  Level Sample
1     0      9
2     1     10
3     2     10
4     3     11
5     4     10
```

The number of samples which were obtained for each participant

```
  Participant Sample
1           A     15
2           B     17
3           C     18
```

Records for samples obtained at treatment levels 0,2 and 4

```
# A tibble: 29 x 3
   sample_id Participant Treatment
   <chr>     <chr>           <dbl>
 1 sample_1  C                   4
 2 sample_3  B                   2
 3 sample_6  A                   4
 4 sample_7  B                   0
 5 sample_9  B                   4
 6 sample_10 C                   0
 7 sample_11 A                   2
 8 sample_14 C                   2
 9 sample_15 A                   0
10 sample_16 C                   4
# i 19 more rows
```

Extract records for samples obtained from participants A and C

```
# A tibble: 33 x 3
   sample_id Participant Treatment
   <chr>     <chr>           <dbl>
 1 sample_1  C                   4
 2 sample_4  C                   1
 3 sample_5  C                   3
 4 sample_6  A                   4
 5 sample_8  A                   1
 6 sample_10 C                   0
 7 sample_11 A                   2
 8 sample_12 A                   3
 9 sample_14 C                   2
10 sample_15 A                   0
# i 23 more rows
```

The number of samples under each participant groups per treatment level.

```
   0 1 2 3 4
A 3 3 3 3 3
B 3 3 4 4 3
C 3 4 3 4 4
```

Section B: Data Exploration 2

AssignmentFile.csv file was imported into R session and the following were the explorations.

The average concentration for each of the concentrations (ie Conc1, Conc2 and Conc3)

```
       Conc1    Conc2   Conc3
Mean 2.23591 2.631801 2.57969
```

The average Concentration for each sample

```
 [1] 3.824863 3.561779 1.499365 2.958203 2.390004 1.868026 2.409794 1.419867
 [9] 3.306693 2.288922 2.056538 1.599757 2.451226 2.901506 3.258827 1.753767
[17] 2.443834 2.657674 1.782624 2.034364 1.998999 2.163061 2.544692 2.452164
[25] 2.092057 3.003388 3.313401 1.765348 1.624170 3.037058 2.448267 2.038269
[33] 2.833067 1.957261 1.668823 2.192923 2.616311 3.646680 2.841996 2.949213
[41] 3.229693 2.329251 3.211285 2.842079 2.998275 2.744949 1.683616 2.672564
[49] 2.434259 2.322598
```

Extended data-frame with a new variable "Concmean" to contain the average Concentration values computed above

```
     samples      Conc1     Conc2      Conc3 Concmean
1   sample_1 5.93804869 2.2877765  3.2487625 3.824863
2   sample_2 3.39846291 4.0508845  3.2359883 3.561779
3   sample_3 1.94647521 1.2918230  1.2597982 1.499365
4   sample_4 1.75043775 4.5088147  2.6153558 2.958203
```

```
5   sample_5 2.90319389 2.6006057  1.6662130 2.390004
6   sample_6 0.74662991 3.5212946  1.3361532 1.868026
7   sample_7 2.61412606 1.7075223  2.9077324 2.409794
8   sample_8 2.86823250 0.6855859  0.7057831 1.419867
9   sample_9 4.25332693 1.5380970  4.1286548 3.306693
10 sample_10 2.53152856 1.4914619  2.8437762 2.288922
11 sample_11 2.02469134 1.0988479  3.0460737 2.056538
12 sample_12 1.68888486 1.5382037  1.5721820 1.599757
13 sample_13 0.07948294 1.8417749  5.4324201 2.451226
14 sample_14 2.35002162 2.8437762  3.5107200 2.901506
15 sample_15 2.39041813 4.7705793  2.6154821 3.258827
16 sample_16 0.61849725 2.2429638  2.3998393 1.753767
17 sample_17 2.67478610 3.2269440  1.4297713 2.443834
18 sample_18 1.31226408 3.2960420  3.3647171 2.657674
19 sample_19 1.51768209 2.1535883  1.6766006 1.782624
20 sample_20 0.40522027 2.5920906  3.1057814 2.034364
21 sample_21 2.11207326 3.4797040  0.4052203 1.998999
22 sample_22 1.41522107 3.1598762  1.9140865 2.163061
23 sample_23 2.48876302 2.8768958  2.2684174 2.544692
24 sample_24 1.88033411 3.1700562  2.3061025 2.452164
25 sample_25 1.60430191 3.2612428  1.4106253 2.092057
26 sample_26 2.74773395 2.8603185  3.4021108 3.003388
27 sample_27 3.10092688 3.8005525  3.0387250 3.313401
28 sample_28 1.50200809 0.8532958  2.9407410 1.765348
29 sample_29 0.96211799 1.7579986  2.1523931 1.624170
30 sample_30 2.14791875 1.9930619  4.9701923 3.037058
31 sample_31 2.29835762 1.8367984  3.2096464 2.448267
32 sample_32 1.79410531 1.8296990  2.4910030 2.038269
33 sample_33 2.38717120 1.7903664  4.3216649 2.833067
34 sample_34 2.18463473 1.4871950  2.1999546 1.957261
35 sample_35 1.97282767 3.3091355 -0.2754937 1.668823
36 sample_36 1.07893276 2.4602617  3.0395745 2.192923
37 sample_37 2.25989664 3.0217533  2.5672832 2.616311
38 sample_38 4.41868056 3.7645799  2.7567790 3.646680
39 sample_39 2.41582396 3.8140188  2.2961438 2.841996
40 sample_40 1.16442633 3.7059658  3.9772471 2.949213
41 sample_41 2.74157875 3.2011782  3.7463223 3.229693
42 sample_42 2.84826466 2.0364330  2.1030553 2.329251
43 sample_43 3.84637857 1.9730300  3.8144462 3.211285
44 sample_44 3.11295419 3.5936637  1.8196195 2.842079
45 sample_45 1.64704709 4.2174458  3.1303307 2.998275
46 sample_46 2.66380292 2.5251532  3.0458905 2.744949
47 sample_47 1.65227360 2.7611225  0.6374504 1.683616
48 sample_48 2.86585757 2.2171705  2.9346651 2.672564
49 sample_49 2.58467415 3.1289636  1.5891398 2.434259
50 sample_50 1.88399429 2.4144523  2.6693480 2.322598
```

The basic statics (min, max, mean, median, quartiles, variance, standard deviation) for all the variables including Concmean

```
     Conc1            Conc2            Conc3            Concmean
 Min.   :0.07948   Min.   :0.6856   Min.   :-0.2755   Min.   :1.420
 1st Qu.:1.64835   1st Qu.:1.8380   1st Qu.: 1.8432   1st Qu.:2.035
 Median :2.22227   Median :2.5963   Median : 2.6424   Median :2.446
```

```
 Mean   :2.23591   Mean   :2.6318   Mean   : 2.5797   Mean   :2.482
 3rd Qu.:2.72488   3rd Qu.:3.2873   3rd Qu.: 3.1898   3rd Qu.:2.937
 Max.   :5.93805   Max.   :4.7706   Max.   : 5.4324   Max.   :3.825
```

```
         Standard_Deviation
Conc1             1.0384617
Conc2             0.9653108
Conc3             1.1169179
Concmean          0.6018165
```

```
          Variance
Conc1    1.0784026
Conc2    0.9318250
Conc3    1.2475056
Concmean 0.3621831
```

Section C : Data Manipulation and Graphics

The new data-frame after removing variables Conc1, Conc2 and Conc3

```
     samples Concmean
1   sample_1 3.824863
2   sample_2 3.561779
3   sample_3 1.499365
4   sample_4 2.958203
5   sample_5 2.390004
6   sample_6 1.868026
7   sample_7 2.409794
8   sample_8 1.419867
9   sample_9 3.306693
10 sample_10 2.288922
11 sample_11 2.056538
12 sample_12 1.599757
13 sample_13 2.451226
14 sample_14 2.901506
15 sample_15 3.258827
16 sample_16 1.753767
17 sample_17 2.443834
18 sample_18 2.657674
19 sample_19 1.782624
20 sample_20 2.034364
21 sample_21 1.998999
22 sample_22 2.163061
23 sample_23 2.544692
24 sample_24 2.452164
25 sample_25 2.092057
26 sample_26 3.003388
27 sample_27 3.313401
28 sample_28 1.765348
29 sample_29 1.624170
30 sample_30 3.037058
31 sample_31 2.448267
32 sample_32 2.038269
```

```
33 sample_33 2.833067
34 sample_34 1.957261
35 sample_35 1.668823
36 sample_36 2.192923
37 sample_37 2.616311
38 sample_38 3.646680
39 sample_39 2.841996
40 sample_40 2.949213
41 sample_41 3.229693
42 sample_42 2.329251
43 sample_43 3.211285
44 sample_44 2.842079
45 sample_45 2.998275
46 sample_46 2.744949
47 sample_47 1.683616
48 sample_48 2.672564
49 sample_49 2.434259
50 sample_50 2.322598
```

Merged information in concentration data-frame with metadata data-frame

```
   sample_id Participant Treatment Concmean
1   sample_1           C         4 3.824863
2   sample_2           B         3 3.561779
3   sample_3           B         2 1.499365
4   sample_4           C         1 2.958203
5   sample_5           C         3 2.390004
6   sample_6           A         4 1.868026
7   sample_7           B         0 2.409794
8   sample_8           A         1 1.419867
9   sample_9           B         4 3.306693
10 sample_10           C         0 2.288922
11 sample_11           A         2 2.056538
12 sample_12           A         3 1.599757
13 sample_13           B         1 2.451226
14 sample_14           C         2 2.901506
15 sample_15           A         0 3.258827
16 sample_16           C         4 1.753767
17 sample_17           B         3 2.443834
18 sample_18           B         2 2.657674
19 sample_19           C         1 1.782624
20 sample_20           C         3 2.034364
21 sample_21           A         4 1.998999
22 sample_22           B         0 2.163061
23 sample_23           A         1 2.544692
24 sample_24           B         4 2.452164
25 sample_25           C         0 2.092057
26 sample_26           A         2 3.003388
27 sample_27           A         3 3.313401
28 sample_28           B         1 1.765348
29 sample_29           C         2 1.624170
30 sample_30           A         0 3.037058
31 sample_31           C         4 2.448267
32 sample_32           B         3 2.038269
```

```
33 sample_33           B         2 2.833067
34 sample_34           C         1 1.957261
35 sample_35           C         3 1.668823
36 sample_36           A         4 2.192923
37 sample_37           B         0 2.616311
38 sample_38           A         1 3.646680
39 sample_39           B         4 2.841996
40 sample_40           C         0 2.949213
41 sample_41           A         2 3.229693
42 sample_42           A         3 2.329251
43 sample_43           B         1 3.211285
44 sample_44           C         2 2.842079
45 sample_45           A         0 2.998275
46 sample_46           C         4 2.744949
47 sample_47           B         3 1.683616
48 sample_48           B         2 2.672564
49 sample_49           C         1 2.434259
50 sample_50           C         3 2.322598
```

Size of the combined data-frame (rows and columns)

```
[1] 50  4
```

Variable "participantsBC" with the extracted records for samples for participants B and C obtained at treatment levels 2,3 and 4.

```
   sample_id Participant Treatment Concmean
1   sample_1           C         4 3.824863
2   sample_2           B         3 3.561779
3   sample_3           B         2 1.499365
4   sample_5           C         3 2.390004
5   sample_9           B         4 3.306693
6  sample_14           C         2 2.901506
7  sample_16           C         4 1.753767
8  sample_17           B         3 2.443834
9  sample_18           B         2 2.657674
10 sample_20           C         3 2.034364
11 sample_24           B         4 2.452164
12 sample_29           C         2 1.624170
13 sample_31           C         4 2.448267
14 sample_32           B         3 2.038269
15 sample_33           B         2 2.833067
16 sample_35           C         3 1.668823
17 sample_39           B         4 2.841996
18 sample_44           C         2 2.842079
19 sample_46           C         4 2.744949
20 sample_47           B         3 1.683616
21 sample_48           B         2 2.672564
22 sample_50           C         3 2.322598
```

Computed average Concentration for participants B and C under treatments 2, 3 and 4.

```
  Participant Treatment Average_Concetration
```

```
1           B         2            2.415668
2           B         3            2.431874
3           B         4            2.866951
4           C         2            2.455918
5           C         3            2.103947
6           C         4            2.692961
```

Changed treatment levels from 0,1,2,3,4 to “very-low”, “low”, “moderate”, “high”, “very-high” respectively on the combined dataset using for loop.

```
 sample_id Participant Treatment Concmean
  sample_1           C very-high 3.824863
  sample_2           B      high 3.561779
  sample_3           B  moderate 1.499365
  sample_4           C       low 2.958203
  sample_5           C      high 2.390004
  sample_6           A very-high 1.868026
  sample_7           B  very-low 2.409794
  sample_8           A       low 1.419867
  sample_9           B very-high 3.306693
 sample_10           C  very-low 2.288922
 sample_11           A  moderate 2.056538
 sample_12           A      high 1.599757
 sample_13           B       low 2.451226
 sample_14           C  moderate 2.901506
 sample_15           A  very-low 3.258827
 sample_16           C very-high 1.753767
 sample_17           B      high 2.443834
 sample_18           B  moderate 2.657674
 sample_19           C       low 1.782624
 sample_20           C      high 2.034364
 sample_21           A very-high 1.998999
 sample_22           B  very-low 2.163061
 sample_23           A       low 2.544692
 sample_24           B very-high 2.452164
 sample_25           C  very-low 2.092057
 sample_26           A  moderate 3.003388
 sample_27           A      high 3.313401
 sample_28           B       low 1.765348
 sample_29           C  moderate 1.624170
 sample_30           A  very-low 3.037058
 sample_31           C very-high 2.448267
 sample_32           B      high 2.038269
 sample_33           B  moderate 2.833067
 sample_34           C       low 1.957261
 sample_35           C      high 1.668823
 sample_36           A very-high 2.192923
 sample_37           B  very-low 2.616311
 sample_38           A       low 3.646680
 sample_39           B very-high 2.841996
 sample_40           C  very-low 2.949213
 sample_41           A  moderate 3.229693
 sample_42           A      high 2.329251
 sample_43           B       low 3.211285
```

```
sample_44           C  moderate 2.842079
sample_45           A  very-low 2.998275
sample_46           C very-high 2.744949
sample_47           B      high 1.683616
sample_48           B  moderate 2.672564
sample_49           C       low 2.434259
sample_50           C      high 2.322598
```

Changed treatment levels from 0,1,2,3,4 to ”very-low”, ”low”, ”moderate”, ”high”, ”very-high” respectively using the created function

```
sample_id Participant Treatment Concmean
 sample_1           C very-high 3.824863
 sample_2           B      high 3.561779
 sample_3           B  moderate 1.499365
 sample_4           C       low 2.958203
 sample_5           C      high 2.390004
 sample_6           A very-high 1.868026
 sample_7           B  very-low 2.409794
 sample_8           A       low 1.419867
 sample_9           B very-high 3.306693
sample_10           C  very-low 2.288922
sample_11           A  moderate 2.056538
sample_12           A      high 1.599757
sample_13           B       low 2.451226
sample_14           C  moderate 2.901506
sample_15           A  very-low 3.258827
sample_16           C very-high 1.753767
sample_17           B      high 2.443834
sample_18           B  moderate 2.657674
sample_19           C       low 1.782624
sample_20           C      high 2.034364
sample_21           A very-high 1.998999
sample_22           B  very-low 2.163061
sample_23           A       low 2.544692
sample_24           B very-high 2.452164
sample_25           C  very-low 2.092057
sample_26           A  moderate 3.003388
sample_27           A      high 3.313401
sample_28           B       low 1.765348
sample_29           C  moderate 1.624170
sample_30           A  very-low 3.037058
sample_31           C very-high 2.448267
sample_32           B      high 2.038269
sample_33           B  moderate 2.833067
sample_34           C       low 1.957261
sample_35           C      high 1.668823
sample_36           A very-high 2.192923
sample_37           B  very-low 2.616311
sample_38           A       low 3.646680
sample_39           B very-high 2.841996
sample_40           C  very-low 2.949213
sample_41           A  moderate 3.229693
sample_42           A      high 2.329251
```

```
sample_43           B       low 3.211285
sample_44           C  moderate 2.842079
sample_45           A  very-low 2.998275
sample_46           C very-high 2.744949
sample_47           B      high 1.683616
sample_48           B  moderate 2.672564
sample_49           C       low 2.434259
sample_50           C      high 2.322598
```

Time taken for for-loop to execute

```
   user  system elapsed
 -0.008   0.000  -0.008
```

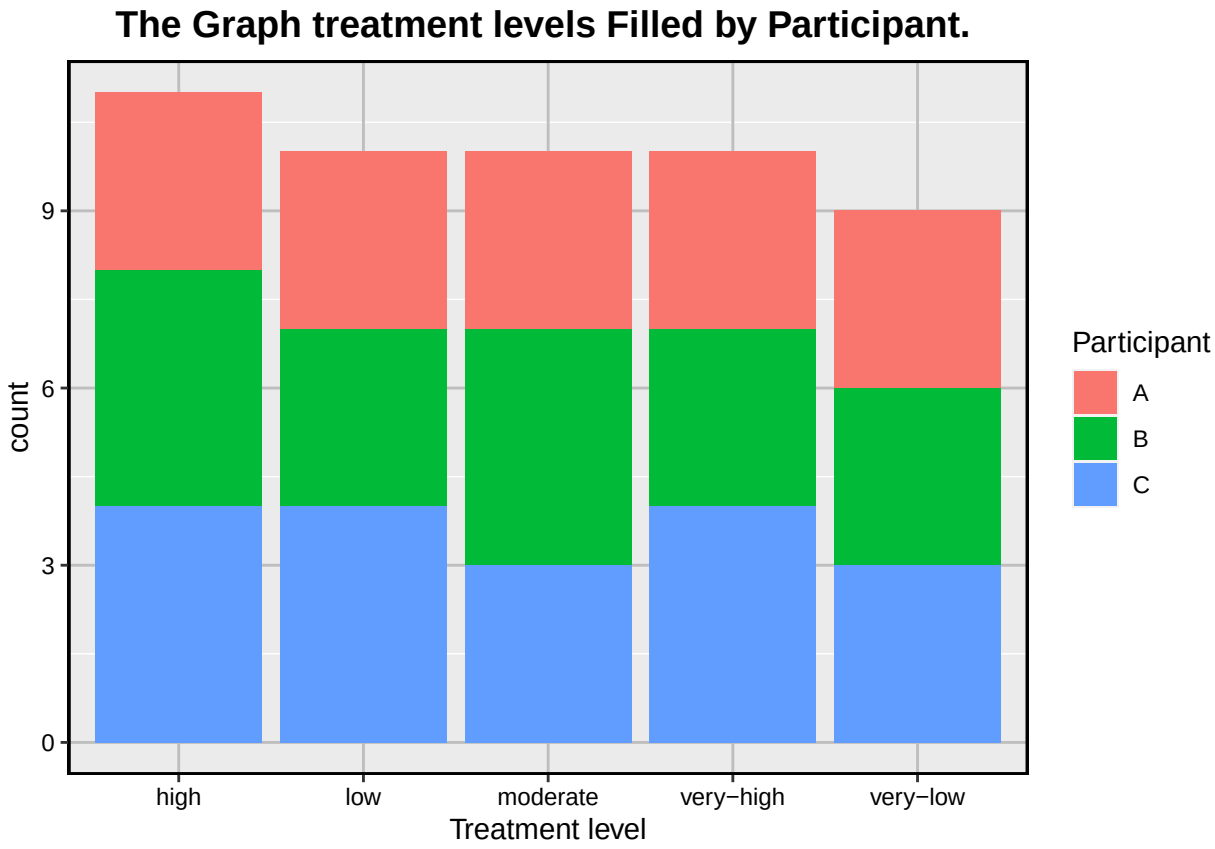
Time taken for sapply() function to execute

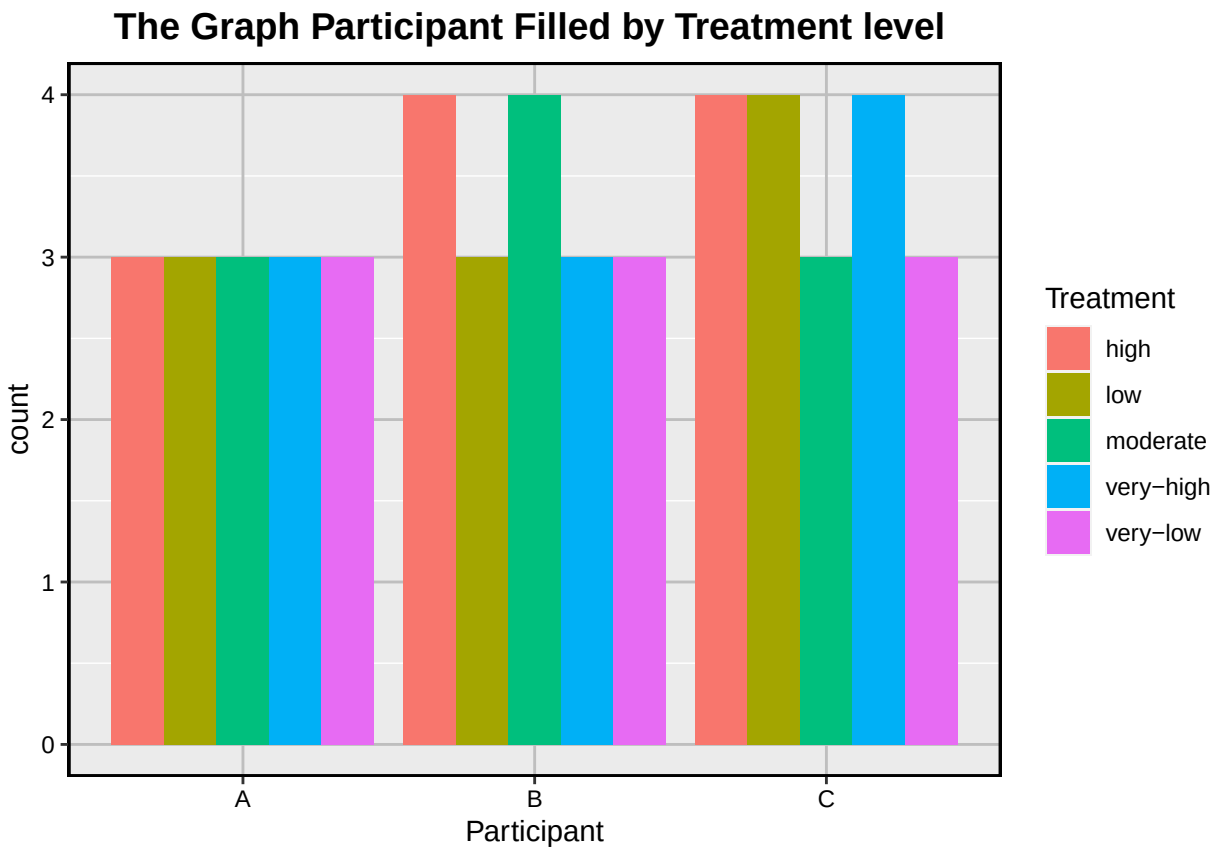
```
   user  system elapsed
  -0.01    0.00   -0.01
```

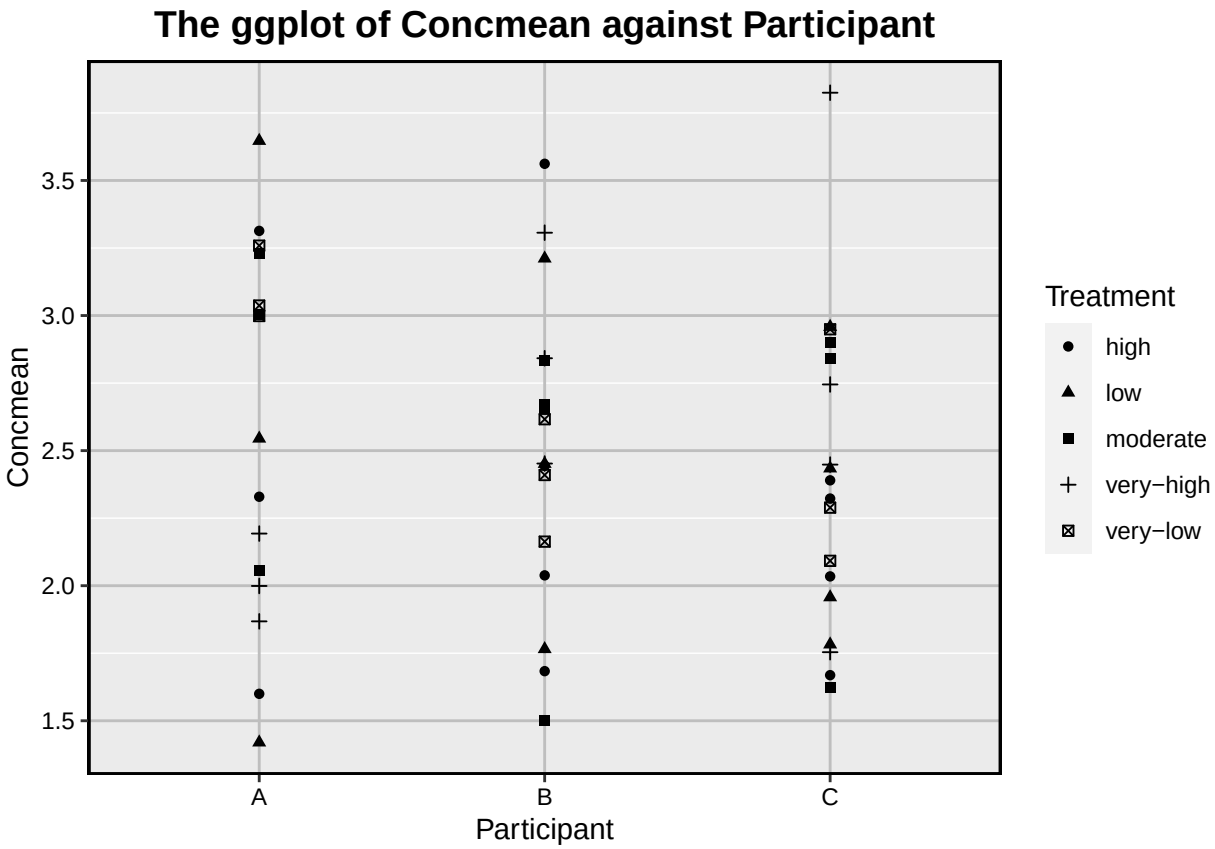
COMMENT: system.time() measures CPU time used by a specific expression or function and the outputs are with variables user, system and elapsed. User Time is the wall clock time. The time that a user experienced. Elapsed Time is the time charged to the CPU(s) for the expression. If elapsed time > user time, this means that the CPU is waiting around for some other operations (may be external) to be done. If elapsed time < user time, this means that the machine has multiple cores and is able to use them. Comparing the above time, sapply() function is faster than for-loop function.

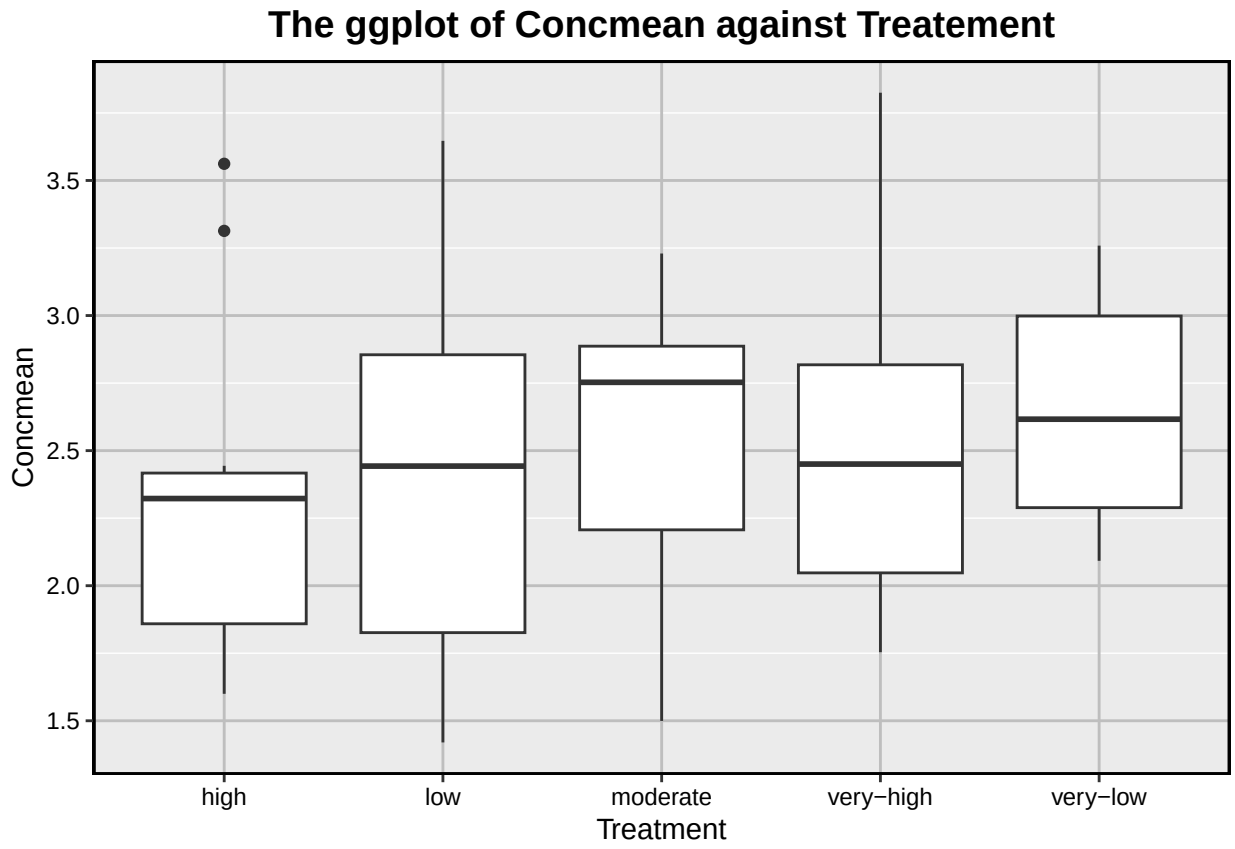
Graphical representations of the combined dataset with the changed treatment levels.

```
Warning: The ‘size‘ argument of ‘element_rect()‘ is deprecated as of ggplot2 3.4.0.
i Please use the ‘linewidth‘ argument instead.
This warning is displayed once every 8 hours.
Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
generated.
```


The Graph treatment levels Filled by Participant.
count
9
6
3
0
high
low
moderate
very−high
very−low
Treatment level
Participant
A
B
C


The Graph Participant Filled by Treatment level
count
4
3
2
1
0
A
B
C
Participant
Treatment
high
low
moderate
very−high
very−low


The ggplot of Concmean against Participant
Concmean
3.5
3.0
2.5
2.0
1.5
A
B
C
Participant
Treatment
high
low
moderate
very−high
very−low



Section D : Classical hypothesis testing and statistical models in R

The statsFile.csv file that contains information about participants involved in a study was read into a data-frame to explore the key predictors of a disease status

The relation of BMI to BP for the all study participants

```
    Pearson's product-moment correlation

data:  stats$BMI and stats$BP
t = 798.75, df = 318, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9996895 0.9998001
sample estimates:
      cor
0.9997509
```

```
    Pearson's product-moment correlation

data:  stats$BMI and stats$BP
t = 798.75, df = 318, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
```

```
95 percent confidence interval:
 0.9996895 0.9998001
sample estimates:
      cor
0.9997509

[1] 0.01065146

[1] 0.9997509

[1] 3.185907

[1] 1.814074


    Welch Two Sample t-test

data:  stats$BMI and stats$BP
t = 121.36, df = 369.13, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.349605 1.394060
sample estimates:
mean of x mean of y
 3.185907  1.814074

[1] 0.03789084

[1] 0.002995712


    F test to compare two variances

data:  stats$BMI and stats$BP
F = 12.648, num df = 319, denom df = 319, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 10.15190 15.75872
sample estimates:
ratio of variances
          12.64836
```

How disease status (phenotype) is related to gender

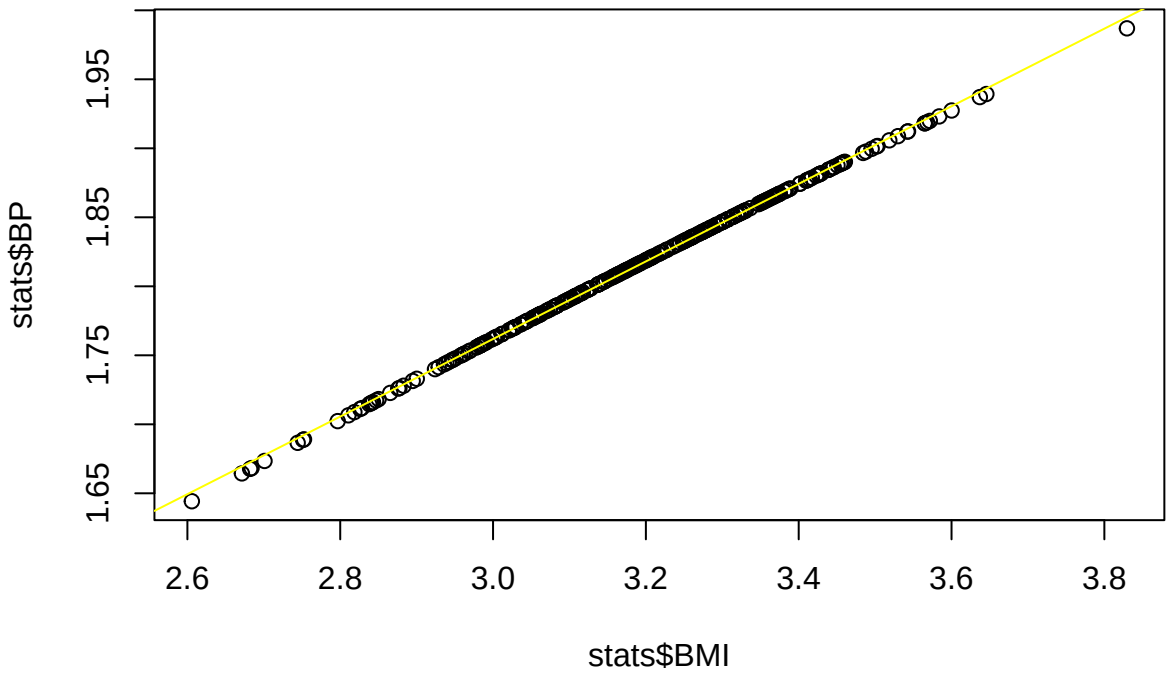
```
    Pearson's Chi-squared test with Yates' continuity correction

data:  table(stats$Status, stats$Gender)
X-squared = 0.86632, df = 1, p-value = 0.352
```

```
   Female Male
0    100   65
1     85   70
```

The constructed model to show the relationship between BMI and BP

```
Call:
lm(formula = BP ~ BMI, data = stats)

Coefficients:
(Intercept)          BMI
     0.9185       0.2811
```

```
Call:
lm(formula = BP ~ BMI, data = stats)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0080845 -0.0001705  0.0004430  0.0007507  0.0008477

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.9184871  0.0011233   817.6   <2e-16 ***
BMI         0.2811090  0.0003519   798.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.001224 on 318 degrees of freedom
Multiple R-squared:  0.9995,    Adjusted R-squared:  0.9995
F-statistic: 6.38e+05 on 1 and 318 DF,  p-value: < 2.2e-16
```



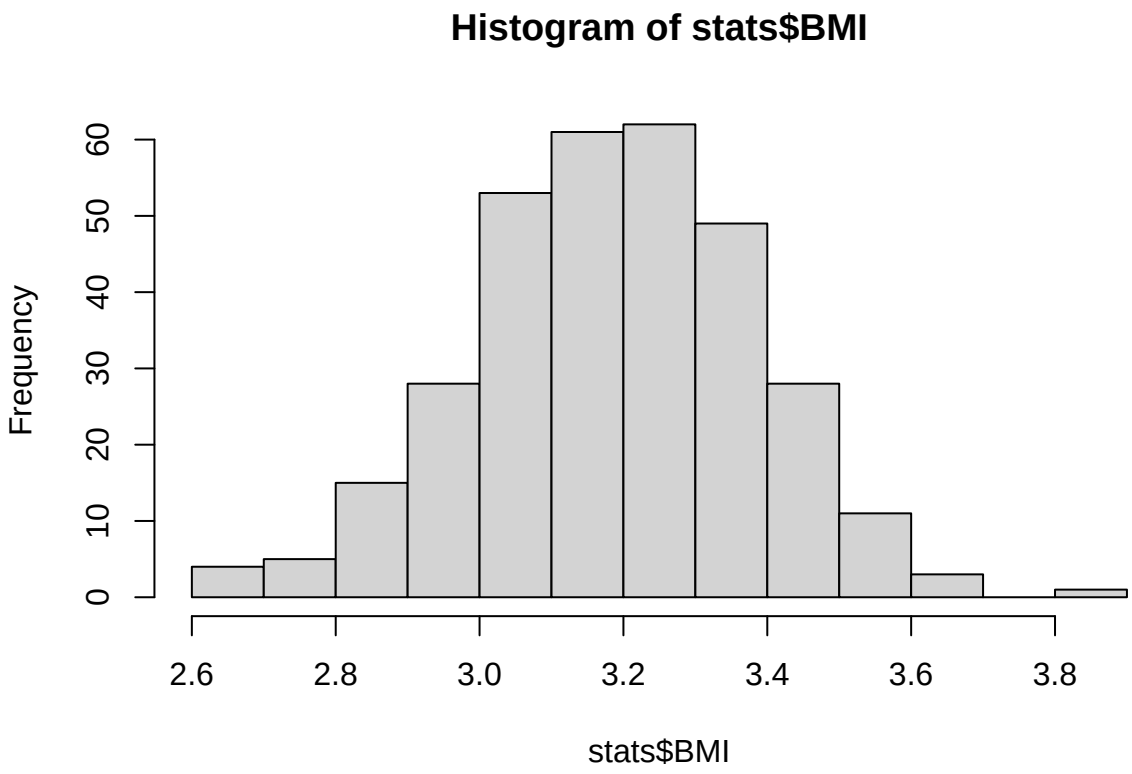
We fitted a linear model (estimated using OLS) to predict BP with BMI (formula: BP ~ BMI). The model explains a statistically significant and substantial proportion of variance (R2 = 1.00, F(1, 318) = 6.38e+05, p < .001, adj. R2 = 1.00). The model's intercept, corresponding to BMI = 0, is at 0.92 (95% CI [0.92, 0.92], t(318) = 817.65, p < .001). Within this model:

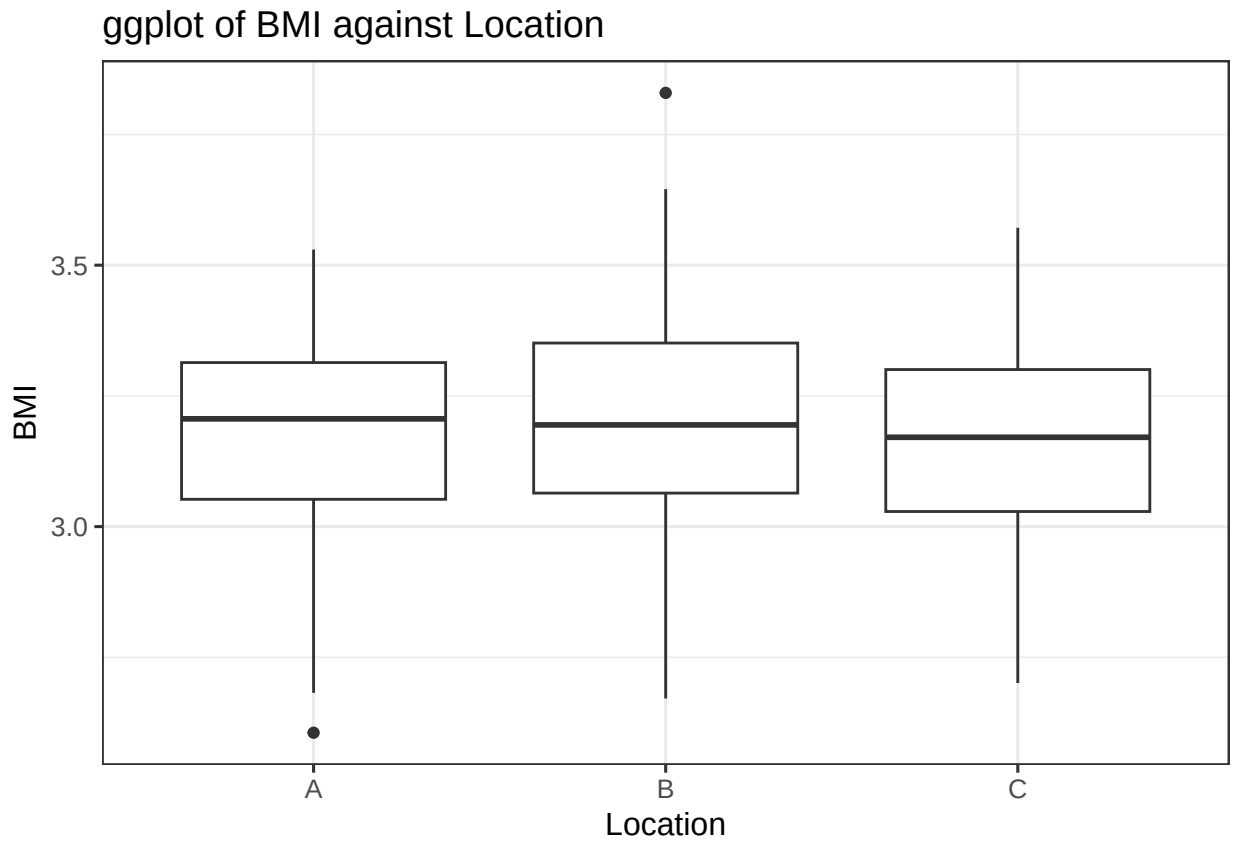
  - The effect of BMI is statistically significant and positive (beta = 0.28, 95% CI [0.28, 0.28], t(318) = 798.75, p < .001; Std. beta = 1.00, 95% CI [1.00, 1.00])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence Intervals (CIs) and p-values were computed using a Wald t-distribution approximation.

The comparison of BMI in the different sampling locations

```
# A tibble: 3 x 4
  Location  mean median variance
  <chr>    <dbl>  <dbl>    <dbl>
1 A         3.18   3.21   0.0369
2 B         3.21   3.19   0.0395
3 C         3.17   3.17   0.0367
```


Histogram of stats$BMI
Frequency
0
10
20
30
40
50
60
2.6
2.8
3.0
3.2
3.4
3.6
3.8
stats$BMI




```
             Df Sum Sq Mean Sq F value Pr(>F)
Location      2  0.097 0.04834   1.278   0.28
Residuals   317 11.991 0.03782

The ANOVA (formula: BMI ~ Location) suggests that:

  - The main effect of Location is statistically not significant and very small
(F(2, 317) = 1.28, p = 0.280; Eta2 = 8.00e-03, 95% CI [0.00, 1.00])

Effect sizes were labelled following Field's (2013) recommendations.
```

How disease status depends on both BMI and BP

```
Call:
glm(formula = Status ~ BMI + BP, family = "binomial", data = stats)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -11.64      84.53  -0.138    0.890
BMI            -4.15      25.88  -0.160    0.873
BP             13.67      92.02   0.149    0.882

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 443.30  on 319  degrees of freedom
Residual deviance: 442.99  on 317  degrees of freedom
AIC: 448.99

Number of Fisher Scoring iterations: 3
```

We fitted a logistic model (estimated using ML) to predict Status with BMI and BP (formula: Status ~ BMI + BP). The model's explanatory power is very weak (Tjur's R2 = 9.60e-04). The model's intercept, corresponding to BMI = 0 and BP = 0, is at -11.64 (95% CI [-180.52, 155.10], p = 0.890). Within this model:

- The effect of BMI is statistically non-significant and negative (beta = -4.15, 95% CI [-55.88, 46.86], p = 0.873; Std. beta = -0.81, 95% CI [-10.88, 9.12])
- The effect of BP is statistically non-significant and positive (beta = 13.67, 95% CI [-167.77, 197.56], p = 0.882; Std. beta = 0.75, 95% CI [-9.18, 10.81])

Standardized parameters were obtained by fitting the model on a standardized version of the dataset. 95% Confidence Intervals (CIs) and p-values were computed using a Wald z-distribution approximation.