CSC 587 HW 2

Daniel R. Getty

2024-02-06

Set r Environment

```
knitr::opts_chunk$set(echo = TRUE, message = TRUE)
# directory
dir <- 'G:\\My Drive\\H Drive\\Course Work\\CERG-Data Science\\CSC_587_Advanced_Data_Mining\\HW\\HW2_Da
# Set the working directory.
setwd(dir)
# Print the working directory.
getwd()
```

```
## [1] "G:/My Drive/H Drive/Course Work/CERG-Data Science/CSC_587_Advanced_Data_Mining/HW/HW2_DataMining
```

```
# load ggplot2 package
library(ggplot2)
# load ggplot2 package
library(ggplot2)
# load dplyr package
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
```

```
# load tidyr package
library(tidyr)
```

Set py Environment

```
import os
import pandas as pd
import numpy as np
import scipy.stats as stats
import math
```

Homework 1

1: Find the distance between objects 1 and 3 by using the formula provided on the slides. Notice that we have mixed type of attributes.

```
# Using Python
# Create a dictionary of data
data1py = {
    'ObjectIdentifier': [1, 2, 3, 4],
    'test1.nominal': ['A', 'B', 'C', 'A'],
    'test2.ordinal': ['excellent', 'fair', 'good', 'excellent'],
    'test3.numeric': [45, 22, 64, 28]
}
# Create a DataFrame from the dictionary

type(data1py)
```

```
## <class 'dict'>
```

```
v1a = data1py['test3.numeric'][0]
v1b = data1py['test3.numeric'][2]
v1c = data1py['ObjectIdentifier']
v1d = data1py['test3.numeric']
print ('v1a =',v1a)
```

```
## v1a = 45
```

```
print ('v1b =',v1b)
```

```
## v1b = 64
```

```
print ('v1c =',v1c)
```

```
## v1c = [1, 2, 3, 4]
```

```
print ('v1d =',v1d)
```

```
## v1d = [45, 22, 64, 28]
```

```
manhattan = abs(v1a - v1b)
euclidian = math.sqrt((v1a - v1b) ** 2)
print('Manhattan =',manhattan)
```

```
## Manhattan = 19
```

```
print('Euclidian =',euclidian)
```

```
## Euclidian = 19.0
```

2: Write a program in any language which can compute Manhattan and Euclidean distances between any two given vectors with any length. You can pass the length to your function, but please don't limit the dimension to 2. You can test your function on vectors you fill in your code without asking user input.

```
# Using Python
def distance(v1, v2):
    # Manhattan distance is taxicab distance, the sum of the absolute differences between the coordinat
    manhatten = sum(abs(a1 - b1) for a1, b1 in zip(v1, v2))
    # Euclidean distance is strait line distance
    euclidian = math.sqrt(sum((a2 - b2) ** 2 for a2, b2 in zip(v1, v2)))
    # Hamming distance is used for categorical data
    hamming = sum(a3 != b3 for a3, b3 in zip(v1, v2))
    # Cosine distance is used to find similarity between data points
    cosine = sum(a4 * b4 for a4, b4 in zip(v1, v2)) / (math.sqrt(sum(a4 ** 2 for a4 in v1)) * math.sqrt
    print("Manhattan distance:", manhatten)
    print("Euclidean distance:", euclidian)
    print("Hamming distance:", hamming)
    print("Cosine distance:", cosine)

# Define two vectors
v1 = v1c
v2 = v1d

# Call  the function
dis = distance(v1, v2)
```

```
## Manhattan distance: 149
## Euclidean distance: 81.44323176298937
## Hamming distance: 4
## Cosine distance: 0.8347166756106098
```

3: In the table below, determine whether passing a class has a dependency on attendance by using Chi-square test. Please refer to the formula in the slides. (For the expected value for each cell, multiply the total counts in the rows and columns of the cell and divide by total count. For example: Expected value for Attended-Pass=33*31/54 = 18.94. You can scan and submit your handwritten calculation)

```
# Using Python
# Create a DataFrame
df = pd.DataFrame({
    'Attended': [25, 6, 31],
    'Skipped': [8, 15, 23],
}, index=['Passed', 'Failed', 'Total'])

# Calculate row and column totals
row_totals = df.loc[:, 'Attended':'Skipped'].sum(axis=1)
col_totals = df.loc['Total', :]
```

```
# Calculate grand total
grand_total = df.loc['Total', 'Attended':'Skipped'].sum()

# Calculate expected values for each cell
expected_values = pd.DataFrame()
for row in ['Passed', 'Failed']:
    for col in ['Attended', 'Skipped']:
        expected_values.loc[row, col] = (row_totals[row] * col_totals[col]) / grand_total
expected_values = round(expected_values,2)

# print the DataFrames
print("The DataFrame is:")
```

```
## The DataFrame is:
```

```
print(df, '\n')
```

```
##         Attended  Skipped
## Passed        25        8
## Failed         6       15
## Total         31       23
```

```
print("The Expected Values are:")
```

```
## The Expected Values are:
```

```
print(expected_values)
```

```
##         Attended  Skipped
## Passed     18.94    14.06
## Failed     12.06     8.94
```

```
# Calculate the chi-square statistic
chi_square = 0
for row in ['Passed', 'Failed']:
    for col in ['Attended', 'Skipped']:
        observed = df.loc[row, col]
        expected = expected_values.loc[row, col]
        chi_square += ((observed - expected) ** 2) / expected
print('\n',"The chi-square statistic is:", chi_square)
```

```
##
##  The chi-square statistic is: 11.703724236949945
```

```
# Calculate the degrees of freedom
degrees_of_freedom = (len(row_totals) - 1) * (len(col_totals) - 1)
print('\n',"The degrees of freedom is:", degrees_of_freedom)
```

```
##
##  The degrees of freedom is: 2
```

```
# chi-square distribution table
# Define the significance level
chialpha = 0.05

# Create a list of degrees of freedom
chidf = list(range(1, 10))

# Calculate critical values for each degree of freedom
critical_values = [stats.chi2.ppf(1 - chialpha, d) for d in chidf]

# Create a DataFrame to display the table
chisqr_ref_tbl = pd.DataFrame({'Degrees_of_Freedom': chidf, 'Critical_Value': critical_values})

# Print the table
print('\n',chisqr_ref_tbl)
```

```
##
##     Degrees_of_Freedom  Critical_Value
## 0                    1        3.841459
## 1                    2        5.991465
## 2                    3        7.814728
## 3                    4        9.487729
## 4                    5       11.070498
## 5                    6       12.591587
## 6                    7       14.067140
## 7                    8       15.507313
## 8                    9       16.918978
```

```
# Calculate the critical value
critical_value = stats.chi2.ppf(1 - chialpha, degrees_of_freedom)
print('\n',"The critical value is:", critical_value)
```

```
##
##  The critical value is: 5.991464547107979
```

```
# Determine whether to reject the null hypothesis
if chi_square > critical_value:
  print('\n',"Reject the null hypothesis")
else:
  print('\n',"Fail to reject the null hypothesis")
```

```
##
##  Reject the null hypothesis
```

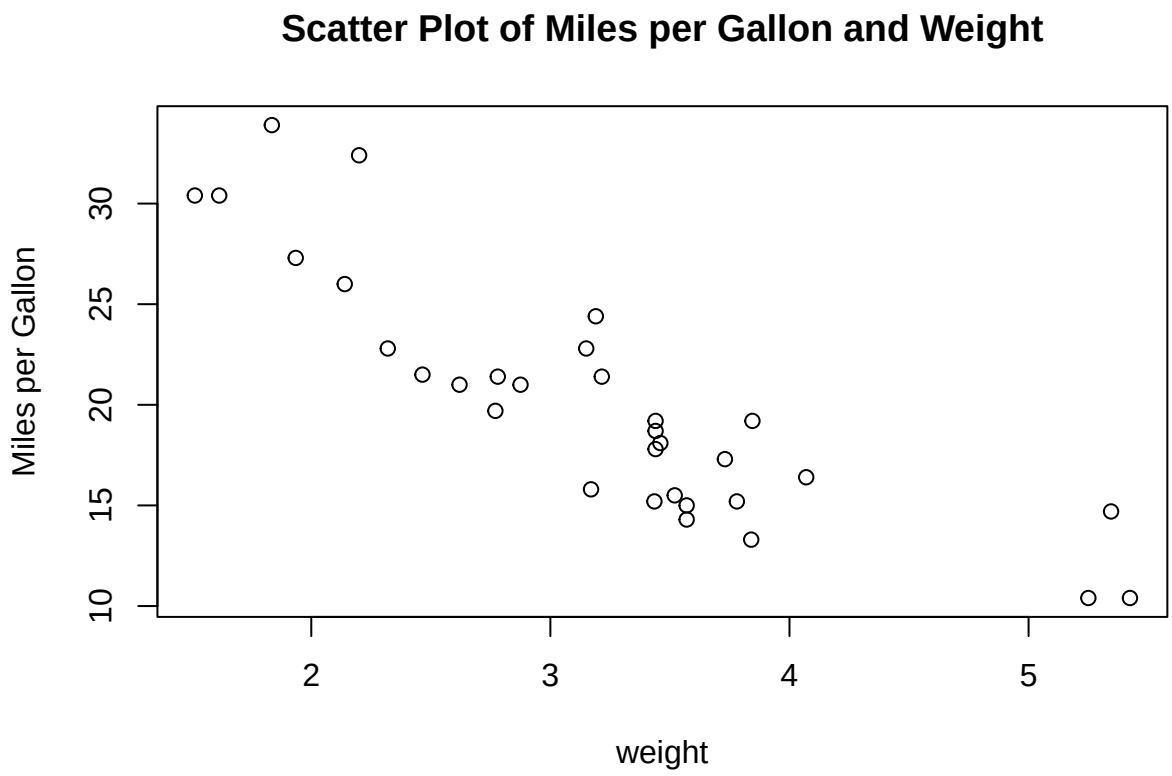
The null hypothesis is rejected. There we can conclude that there is a statistically significant correlation between attendance and the likelihood of passing a class.

4: In R, there is a built-in data frame called mtcars. Please calculate the correlation between mpg and wt attributes of mtcars by using cor() function. Then generate scatter plot based on these two attributes. Your scatter plot should be like the one below. You don't need to submit the image, but R script should be submitted

```r
# Using R
# Load the mtcars data
data(mtcars)
# Calculate the correlation between mpg and wt
cor(mtcars$mpg, mtcars$wt)
```

```
## [1] -0.8676594
```

```r
# Generate scatter plot
plot(mtcars$wt, mtcars$mpg, xlab='weight', ylab='Miles per Gallon', main='Scatter Plot of Miles per Gal
```



5: Grad Students Only Write an R or Python script which removes or drops the columns which have more than 75% missing values. Then it should replace the missing values in the remaining columns with the median value of the 1existing values of that particular column. Download metabolite.csv from Google Drive and use this data set to test your code. Please check the end of this document for some useful R examples and hints.

```r
#` Using R
# Load the metabolite data
data_file <- file.path('metabolite.csv')
# Build data frame from the data set.
metabolite <-read.csv(data_file, header = TRUE, sep = ',')
# Print the data frame.
#glimpse(metabolite)
head(metabolite)
```

```
##       Label  Phe Pro Ser Thr ADMA alpha.AAA c4.OH.Pro Carnosine Creatinine
## 1 Alzheimer 72.8 166 170 282 1.15     0.760     0.236     1.270       49.9
## 2 Alzheimer 93.4 138 142 217 1.05     0.929     0.189     1.350       48.8
## 3 Alzheimer 68.6 161 158 208 1.00     0.620     0.198     0.998       30.4
## 4 Alzheimer 94.1 129 162 201 1.10     0.795        NA     0.675       80.1
## 5 Alzheimer 79.8 126 115 199 1.24     1.360        NA     1.280       60.5
## 6 Alzheimer 82.5 167 173 333 1.35     1.150        NA     1.010       24.0
##    DOPA Dopamine Histamine Kynurenine Met.SO Nitro.Tyr PEA Putrescine Sarcosine
## 1 0.265    0.233     0.225       5.21  0.526     0.027  NA      0.068      17.8
## 2 0.252       NA     0.211       5.44  0.387        NA  NA      0.087      20.2
## 3 0.268       NA     0.217       5.20  0.651        NA  NA      0.260      14.4
## 4 0.264    0.234     0.209       5.80  0.389        NA  NA      0.110      18.7
## 5 0.271    0.231     0.210       4.46  0.466        NA  NA      0.118      22.5
## 6 0.275       NA     0.212       7.01  0.417        NA  NA      0.262      30.8
##   Serotonin Spermidine Spermine t4.OH.Pro Taurine SDMA   C0   C10 C10.1 C10.2
## 1     0.147      0.188       NA      24.0     125 1.13 18.2 0.059 0.312 0.038
## 2     0.231      0.233       NA      29.3     120 1.65 17.0 0.051 0.288 0.039
## 3     0.196      0.384       NA      20.9     139 1.57 12.6 0.083 0.357 0.054
## 4     0.255      0.353       NA      23.1     159 1.34 23.5 0.071 0.317 0.040
## 5     0.390      0.473       NA      26.9     149 1.24 13.6 0.139 0.472 0.074
## 6     0.140      0.856     1.28      26.0     379 1.44 26.7 0.058 0.238 0.042
##     C12 C12.DC C12.1   C14 C14.1 C14.1.OH C14.2 C14.2.OH   C16 C16.OH C16.1
## 1 0.030  0.042 0.290 0.023 0.019    0.008 0.008    0.006 0.046  0.008 0.009
## 2 0.038  0.038 0.265 0.026 0.017    0.008 0.009    0.009 0.070  0.009 0.013
## 3 0.032  0.048 0.302 0.021 0.031    0.010 0.010    0.009 0.076  0.011 0.019
## 4 0.045  0.048 0.275 0.026 0.028    0.010 0.013    0.011 0.074  0.011 0.015
## 5 0.056  0.079 0.394 0.034 0.043    0.016 0.025    0.017 0.062     NA 0.024
## 6 0.039  0.035 0.196 0.029 0.023    0.009 0.010    0.007 0.081  0.006 0.012
##   C16.1.OH C16.2 C16.2.OH   C18 C18.1 C18.1.OH C18.2   C2    C3 C3.OH  C3.1
## 1    0.007 0.005    0.013 0.013 0.024    0.003 0.016 1.97 0.354 0.008 0.015
## 2    0.006 0.006    0.012 0.014 0.025    0.003 0.028 1.95 0.184 0.009 0.013
## 3    0.010 0.005    0.013 0.016 0.025       NA 0.018 1.70 0.371    NA 0.012
## 4    0.008 0.006    0.009 0.020 0.035    0.004 0.033 2.10 0.278 0.010 0.017
## 5    0.014 0.012    0.025 0.031 0.034    0.012 0.017 5.62 0.436 0.029 0.035
## 6    0.005 0.007    0.015 0.017 0.035    0.004 0.029 3.49 0.461 0.008 0.014
##     C4 C3.DC..C4.OH.  C4.1    C5 C5.M.DC C5.OH..C3.DC.M.  C5.1 C5.1.DC
## 1 0.082         0.045 0.025 0.094   0.023           0.026 0.030   0.020
## 2 0.108         0.080 0.025 0.077   0.032           0.026 0.024   0.021
## 3 0.057         0.035 0.039 0.096   0.045           0.024 0.037   0.018
## 4 0.110         0.077 0.031 0.145   0.034           0.041 0.035   0.016
## 5 0.106         0.099 0.069 0.141   0.094           0.058 0.073   0.049
## 6 0.123         0.068 0.026 0.090   0.019           0.037 0.022   0.016
##   C6..C4.1.DC. C5.DC..C6.OH.  C6.1 C7.DC    C8    C9 lysoPC.a.C14.0
## 1        0.022         0.014 0.018 0.011 0.062 0.016           2.23
## 2        0.030         0.018 0.015 0.010 0.058 0.014           1.97
## 3        0.022         0.029 0.031 0.021 0.090 0.017           2.12
## 4        0.029         0.016 0.027 0.017 0.091 0.018           2.19
## 5        0.052         0.040 0.040 0.036 0.192 0.041           1.88
## 6        0.063         0.016 0.019 0.014 0.073 0.014           2.11
##   lysoPC.a.C16.0 lysoPC.a.C16.1 lysoPC.a.C17.0 lysoPC.a.C18.0 lysoPC.a.C18.1
## 1           37.9           2.66          0.446           9.00           8.58
## 2           22.1           1.31          0.270           5.35           3.94
## 3           33.7           2.53          0.399           7.51           7.73
## 4           32.8           2.39          0.323           7.21           7.22
```

```
## 5           24.5           1.27          0.382           6.66           5.39
## 6           29.1           2.09          0.348           5.84           6.30
##   lysoPC.a.C18.2 lysoPC.a.C20.3 lysoPC.a.C20.4 lysoPC.a.C24.0 lysoPC.a.C26.0
## 1           7.27          1.830           8.25          0.079          0.113
## 2           4.42          0.958           4.60          0.059          0.066
## 3           8.02          2.050           9.84          0.075          0.126
## 4           7.62          1.640           6.75          0.066          0.086
## 5           3.60          0.970           6.26          0.084          0.118
## 6           8.10          1.970           7.04          0.083          0.112
##   lysoPC.a.C26.1 lysoPC.a.C28.0 lysoPC.a.C28.1 PC.aa.C24.0 PC.aa.C26.0
## 1          0.053          0.108          0.072       0.082       0.438
## 2          0.042          0.076          0.058       0.065       0.409
## 3          0.049          0.078          0.092       0.099       0.458
## 4          0.045          0.076          0.076       0.076       0.486
## 5          0.053          0.092          0.072       0.069       0.401
## 6          0.050          0.099          0.083       0.073       0.450
##   PC.aa.C28.1 PC.aa.C30.0 PC.aa.C32.0 PC.aa.C32.1 PC.aa.C32.2 PC.aa.C32.3
## 1       0.571        2.35        11.4        9.22          NA       0.092
## 2       0.521        1.99        12.7        5.40          NA       0.067
## 3       0.605        2.69        16.6       11.60          NA       0.105
## 4       0.685        3.33        18.6       13.30       0.053       0.079
## 5       0.513        1.78        13.8        5.03          NA       0.102
## 6       0.620        2.61        14.7        8.98          NA       0.107
##   PC.aa.C34.1 PC.aa.C34.2 PC.aa.C34.3 PC.aa.C34.4 PC.aa.C36.0 PC.aa.C36.1
## 1       109.0        71.0       1.430       0.200        2.38        21.7
## 2        64.2        60.5       0.879       0.127        2.05        14.3
## 3       108.0        83.1       1.930       0.210        2.30        19.9
## 4       106.0        93.6       1.590       0.190        2.57        20.9
## 5        83.4        35.9       0.709       0.135        1.83        20.5
## 6        90.2        85.6       1.790       0.213        2.48        15.5
##   PC.aa.C36.2 PC.aa.C36.3 PC.aa.C36.4 PC.aa.C36.5 PC.aa.C36.6 PC.aa.C38.0
## 1        42.4        42.7       120.0        1.86       0.084       1.230
## 2        35.6        24.3        83.7        1.05       0.046       0.946
## 3        44.9        43.9       146.0        2.09       0.057       1.210
## 4        48.8        41.2       122.0        1.76       0.070       1.160
## 5        28.5        21.9        98.1        1.70       0.048       1.100
## 6        43.2        46.0       114.0        3.47       0.103       1.390
##   PC.aa.C38.3 PC.aa.C38.4 PC.aa.C38.5 PC.aa.C38.6 PC.aa.C40.1 PC.aa.C40.2
## 1        32.1        95.1       16.80        41.6       0.195       0.074
## 2        21.9        78.9        9.91        25.1       0.211       0.057
## 3        34.5       107.0       17.50        36.6       0.212       0.118
## 4        28.7        92.7       14.30        29.9       0.220       0.097
## 5        23.3       101.0       13.80        36.2       0.165       0.044
## 6        28.9        78.0       13.10        48.4       0.205       0.120
##   PC.aa.C40.3 PC.aa.C40.4 PC.aa.C40.5 PC.aa.C40.6 PC.aa.C42.0 PC.aa.C42.1
## 1       0.491        3.48        5.66        21.8       0.364       0.226
## 2       0.358        3.39        4.08        14.2       0.419       0.216
## 3       0.395        3.56        5.34        16.7       0.476       0.281
## 4       0.433        3.59        5.06        14.0       0.427       0.223
## 5       0.525        3.37        5.29        22.5       0.125       0.095
## 6       0.346        2.63        3.25        18.9       0.451       0.233
##   PC.aa.C42.2 PC.aa.C42.4 PC.aa.C42.5 PC.aa.C42.6 PC.ae.C30.0 PC.ae.C30.1
## 1       0.108       0.272       0.272       0.291       0.173       0.027
## 2       0.109       0.336       0.317       0.248       0.147       0.024
```

```
## 3       0.118       0.300       0.206       0.267       0.209       0.046
## 4       0.119       0.268       0.267       0.254       0.223       0.049
## 5       0.083       0.206       0.205       0.280       0.095       0.082
## 6       0.135       0.228       0.254       0.271       0.221       0.039
##   PC.ae.C30.2 PC.ae.C32.1 PC.ae.C32.2 PC.ae.C34.0 PC.ae.C34.1 PC.ae.C34.2
## 1       0.022        1.65       0.371       0.880        3.66        2.48
## 2       0.020        2.01       0.360       0.763        2.68        2.32
## 3       0.030        2.40       0.477       0.938        4.04        2.95
## 4       0.023        2.47       0.459       0.964        4.06        3.09
## 5       0.023        1.72       0.316       1.060        3.28        1.70
## 6       0.029        2.01       0.397       0.920        3.26        2.58
##   PC.ae.C34.3 PC.ae.C36.0 PC.ae.C36.1 PC.ae.C36.2 PC.ae.C36.3 PC.ae.C36.4
## 1       0.813       0.498        5.64        1.90       1.170        6.96
## 2       0.905       0.398        3.89        1.54       0.873        6.40
## 3       1.030       0.554        5.95        2.29       1.240        9.05
## 4       1.020       0.552        4.75        2.01       1.350        8.36
## 5       0.722       0.553        5.95        1.47       0.760        4.78
## 6       1.000       0.443        4.95        2.05       1.170        7.04
##   PC.ae.C36.5 PC.ae.C38.0 PC.ae.C38.1 PC.ae.C38.2 PC.ae.C38.3 PC.ae.C38.4
## 1        4.79       0.474       0.287       0.538        2.66        6.33
## 2        5.36       0.325          NA       0.127        1.80        5.37
## 3        6.63       0.478       0.285       0.154        2.87        7.06
## 4        5.97       0.397       0.022       0.144        1.97        5.99
## 5        4.00       0.430       0.271       0.246        1.80        5.45
## 6        4.47       0.590          NA       0.312        2.46        5.55
##   PC.ae.C38.5 PC.ae.C38.6 PC.ae.C40.1 PC.ae.C40.2 PC.ae.C40.3 PC.ae.C40.4
## 1        5.51        1.95       0.574       0.575       0.940        1.76
## 2        4.49        1.63       0.281       0.491       0.702        1.43
## 3        5.64        1.98       0.759       0.654       0.817        1.51
## 4        5.63        1.97       0.425       0.540       0.742        1.45
## 5        4.34        1.51       0.430       0.432       0.632        1.10
## 6        4.60        1.80       0.481       0.598       0.826        1.25
##   PC.ae.C40.5 PC.ae.C40.6 PC.ae.C42.0 PC.ae.C42.1 PC.ae.C42.2 PC.ae.C42.3
## 1        1.77        1.59       0.629       0.316       0.192       0.277
## 2        1.55        1.20       0.616       0.260       0.157       0.200
## 3        1.64        1.49       0.686       0.356       0.241       0.288
## 4        1.62        1.25       0.637       0.299       0.159       0.208
## 5        1.25        1.47       0.660       0.355       0.138       0.174
## 6        1.38        1.61       0.669       0.265       0.195       0.253
##   PC.ae.C42.4 PC.ae.C42.5 PC.ae.C44.3 PC.ae.C44.4 PC.ae.C44.5 PC.ae.C44.6
## 1       0.264       0.888       0.065       0.168       0.536       0.494
## 2       0.311       0.840       0.071       0.220       0.470       0.515
## 3       0.319       0.957       0.065       0.228       0.565       0.603
## 4       0.392       0.863       0.069       0.237       0.517       0.611
## 5       0.162       0.513       0.081       0.154       0.178       0.134
## 6       0.316       0.814       0.085       0.232       0.554       0.539
##   SM..OH..C14.1 SM..OH..C16.1 SM..OH..C22.1 SM..OH..C22.2 SM..OH..C24.1
## 1         1.420          1.33          2.07          1.86         0.597
## 2         1.390          1.25          2.47          2.20         0.640
## 3         1.840          1.58          2.69          2.63         0.665
## 4         1.720          1.48          2.97          2.84         0.682
## 5         0.987          1.48          1.96          1.74         0.478
## 6         1.320          1.12          2.51          2.16         0.640
##   SM.C16.0 SM.C16.1 SM.C18.0 SM.C18.1 SM.C20.2 SM.C24.0 SM.C24.1 SM.C26.0
```

```
## 1     44.9     7.99     14.5    10.40    0.290    12.20     27.3    0.147
## 2     42.1     6.88     12.7     8.52    0.211    10.40     25.6    0.130
## 3     44.8     8.91     14.6    11.60    0.304    11.50     28.8    0.163
## 4     52.4     8.61     17.2    11.50    0.261    11.80     27.9    0.138
## 5     40.6     5.86     13.0     8.34    0.196     9.29     20.5    0.111
## 6     42.6     8.49     13.0    10.60    0.270     9.58     23.7    0.135
##   SM.C26.1 H1_1   H1 Urea_N L.Arginine_N L.Leucine_N EDTAca_N
## 1    0.337 3356 3356     NA           NA          NA       NA
## 2    0.317 2509 2509  201.9         22.5        35.3      2.0
## 3    0.364 2661 2661  193.3         21.0        25.4      1.8
## 4    0.353 2652 2652  500.8         16.0        27.1      2.5
## 5    0.283 2258 2258  132.5         13.2        57.9      2.5
## 6    0.316 3031 3031  193.3         32.2        26.5      0.0
##   X2.Hydroxybutyrate X3.Hydroxybutyrate Acetate Acetoacetate Acetone Betaine
## 1                 NA                 NA      NA           NA      NA      NA
## 2              12.40                8.5    13.2          5.7     5.1    22.0
## 3              11.33               11.7     5.8          9.3     5.6    19.1
## 4              12.70                7.2     9.8          4.8     4.0    13.9
## 5              35.20               44.7    20.2         18.9    18.9    33.9
## 6              17.20               16.0    23.6          7.8     5.5    16.9
##   Carnitine Choline Creatine Dimethyl.sulfone Ethanol Formate Glucose Glycerol
## 1        NA      NA       NA               NA      NA      NA      NA       NA
## 2       8.7    14.2     14.5              4.7    16.6    24.6  1489.7    324.6
## 3      15.3    14.5     17.8              2.1     8.1    27.4  1343.9    201.3
## 4       7.7    11.8     14.7              1.3     6.4    14.4   629.5    322.0
## 5      18.5    27.7     35.4              5.5    13.0    40.0  1618.0    271.6
## 6      16.7    25.9     18.6              3.4     5.0    35.5  1791.8    274.2
##   Hypoxanthine Isobutyrate Isopropanol Lactate Malonate
## 1           NA          NA          NA      NA       NA
## 2          6.3         3.6         1.9  1171.6     10.4
## 3          6.0         2.5         2.5  1938.1     13.1
## 4          8.6         2.5         4.4  1037.7      7.6
## 5          0.0         6.1        11.2  2199.9     11.7
## 6          8.8         2.3         2.4  1486.7     11.8
```

```
# Remove columns with more than 75% missing values by keeping the columns with less than 75% missing va
clean_metabolite <- metabolite[, colSums(is.na(metabolite)) <= 0.75 * nrow(metabolite)]
head(clean_metabolite)
```

```
##       Label  Phe Pro Ser Thr ADMA alpha.AAA c4.OH.Pro Carnosine Creatinine
## 1 Alzheimer 72.8 166 170 282 1.15     0.760     0.236     1.270       49.9
## 2 Alzheimer 93.4 138 142 217 1.05     0.929     0.189     1.350       48.8
## 3 Alzheimer 68.6 161 158 208 1.00     0.620     0.198     0.998       30.4
## 4 Alzheimer 94.1 129 162 201 1.10     0.795        NA     0.675       80.1
## 5 Alzheimer 79.8 126 115 199 1.24     1.360        NA     1.280       60.5
## 6 Alzheimer 82.5 167 173 333 1.35     1.150        NA     1.010       24.0
##    DOPA Dopamine Histamine Kynurenine Met.SO Putrescine Sarcosine Serotonin
## 1 0.265    0.233     0.225       5.21  0.526      0.068      17.8     0.147
## 2 0.252       NA     0.211       5.44  0.387      0.087      20.2     0.231
## 3 0.268       NA     0.217       5.20  0.651      0.260      14.4     0.196
## 4 0.264    0.234     0.209       5.80  0.389      0.110      18.7     0.255
## 5 0.271    0.231     0.210       4.46  0.466      0.118      22.5     0.390
## 6 0.275       NA     0.212       7.01  0.417      0.262      30.8     0.140
##   Spermidine t4.OH.Pro Taurine SDMA   C0   C10 C10.1 C10.2   C12 C12.DC C12.1
```

```
## 1      0.188      24.0     125 1.13 18.2 0.059 0.312 0.038 0.030  0.042 0.290
## 2      0.233      29.3     120 1.65 17.0 0.051 0.288 0.039 0.038  0.038 0.265
## 3      0.384      20.9     139 1.57 12.6 0.083 0.357 0.054 0.032  0.048 0.302
## 4      0.353      23.1     159 1.34 23.5 0.071 0.317 0.040 0.045  0.048 0.275
## 5      0.473      26.9     149 1.24 13.6 0.139 0.472 0.074 0.056  0.079 0.394
## 6      0.856      26.0     379 1.44 26.7 0.058 0.238 0.042 0.039  0.035 0.196
##     C14 C14.1 C14.1.OH C14.2 C14.2.OH   C16 C16.OH C16.1 C16.1.OH C16.2
## 1 0.023 0.019    0.008 0.008    0.006 0.046  0.008 0.009    0.007 0.005
## 2 0.026 0.017    0.008 0.009    0.009 0.070  0.009 0.013    0.006 0.006
## 3 0.021 0.031    0.010 0.010    0.009 0.076  0.011 0.019    0.010 0.005
## 4 0.026 0.028    0.010 0.013    0.011 0.074  0.011 0.015    0.008 0.006
## 5 0.034 0.043    0.016 0.025    0.017 0.062     NA 0.024    0.014 0.012
## 6 0.029 0.023    0.009 0.010    0.007 0.081  0.006 0.012    0.005 0.007
##   C16.2.OH   C18 C18.1 C18.1.OH C18.2   C2    C3 C3.OH  C3.1    C4
## 1    0.013 0.013 0.024    0.003 0.016 1.97 0.354 0.008 0.015 0.082
## 2    0.012 0.014 0.025    0.003 0.028 1.95 0.184 0.009 0.013 0.108
## 3    0.013 0.016 0.025       NA 0.018 1.70 0.371    NA 0.012 0.057
## 4    0.009 0.020 0.035    0.004 0.033 2.10 0.278 0.010 0.017 0.110
## 5    0.025 0.031 0.034    0.012 0.017 5.62 0.436 0.029 0.035 0.106
## 6    0.015 0.017 0.035    0.004 0.029 3.49 0.461 0.008 0.014 0.123
##   C3.DC..C4.OH.  C4.1    C5 C5.M.DC C5.OH..C3.DC.M.  C5.1 C5.1.DC C6..C4.1.DC.
## 1         0.045 0.025 0.094   0.023           0.026 0.030   0.020        0.022
## 2         0.080 0.025 0.077   0.032           0.026 0.024   0.021        0.030
## 3         0.035 0.039 0.096   0.045           0.024 0.037   0.018        0.022
## 4         0.077 0.031 0.145   0.034           0.041 0.035   0.016        0.029
## 5         0.099 0.069 0.141   0.094           0.058 0.073   0.049        0.052
## 6         0.068 0.026 0.090   0.019           0.037 0.022   0.016        0.063
##   C5.DC..C6.OH.  C6.1 C7.DC    C8    C9 lysoPC.a.C14.0 lysoPC.a.C16.0
## 1         0.014 0.018 0.011 0.062 0.016           2.23           37.9
## 2         0.018 0.015 0.010 0.058 0.014           1.97           22.1
## 3         0.029 0.031 0.021 0.090 0.017           2.12           33.7
## 4         0.016 0.027 0.017 0.091 0.018           2.19           32.8
## 5         0.040 0.040 0.036 0.192 0.041           1.88           24.5
## 6         0.016 0.019 0.014 0.073 0.014           2.11           29.1
##   lysoPC.a.C16.1 lysoPC.a.C17.0 lysoPC.a.C18.0 lysoPC.a.C18.1 lysoPC.a.C18.2
## 1           2.66          0.446           9.00           8.58           7.27
## 2           1.31          0.270           5.35           3.94           4.42
## 3           2.53          0.399           7.51           7.73           8.02
## 4           2.39          0.323           7.21           7.22           7.62
## 5           1.27          0.382           6.66           5.39           3.60
## 6           2.09          0.348           5.84           6.30           8.10
##   lysoPC.a.C20.3 lysoPC.a.C20.4 lysoPC.a.C24.0 lysoPC.a.C26.0 lysoPC.a.C26.1
## 1          1.830           8.25          0.079          0.113          0.053
## 2          0.958           4.60          0.059          0.066          0.042
## 3          2.050           9.84          0.075          0.126          0.049
## 4          1.640           6.75          0.066          0.086          0.045
## 5          0.970           6.26          0.084          0.118          0.053
## 6          1.970           7.04          0.083          0.112          0.050
##   lysoPC.a.C28.0 lysoPC.a.C28.1 PC.aa.C24.0 PC.aa.C26.0 PC.aa.C28.1 PC.aa.C30.0
## 1          0.108          0.072       0.082       0.438       0.571        2.35
## 2          0.076          0.058       0.065       0.409       0.521        1.99
## 3          0.078          0.092       0.099       0.458       0.605        2.69
## 4          0.076          0.076       0.076       0.486       0.685        3.33
## 5          0.092          0.072       0.069       0.401       0.513        1.78
```

```
## 6          0.099          0.083       0.073       0.450       0.620        2.61
##   PC.aa.C32.0 PC.aa.C32.1 PC.aa.C32.2 PC.aa.C32.3 PC.aa.C34.1 PC.aa.C34.2
## 1        11.4        9.22          NA       0.092       109.0        71.0
## 2        12.7        5.40          NA       0.067        64.2        60.5
## 3        16.6       11.60          NA       0.105       108.0        83.1
## 4        18.6       13.30       0.053       0.079       106.0        93.6
## 5        13.8        5.03          NA       0.102        83.4        35.9
## 6        14.7        8.98          NA       0.107        90.2        85.6
##   PC.aa.C34.3 PC.aa.C34.4 PC.aa.C36.0 PC.aa.C36.1 PC.aa.C36.2 PC.aa.C36.3
## 1       1.430       0.200        2.38        21.7        42.4        42.7
## 2       0.879       0.127        2.05        14.3        35.6        24.3
## 3       1.930       0.210        2.30        19.9        44.9        43.9
## 4       1.590       0.190        2.57        20.9        48.8        41.2
## 5       0.709       0.135        1.83        20.5        28.5        21.9
## 6       1.790       0.213        2.48        15.5        43.2        46.0
##   PC.aa.C36.4 PC.aa.C36.5 PC.aa.C36.6 PC.aa.C38.0 PC.aa.C38.3 PC.aa.C38.4
## 1       120.0        1.86       0.084       1.230        32.1        95.1
## 2        83.7        1.05       0.046       0.946        21.9        78.9
## 3       146.0        2.09       0.057       1.210        34.5       107.0
## 4       122.0        1.76       0.070       1.160        28.7        92.7
## 5        98.1        1.70       0.048       1.100        23.3       101.0
## 6       114.0        3.47       0.103       1.390        28.9        78.0
##   PC.aa.C38.5 PC.aa.C38.6 PC.aa.C40.1 PC.aa.C40.2 PC.aa.C40.3 PC.aa.C40.4
## 1       16.80        41.6       0.195       0.074       0.491       3.48
## 2        9.91        25.1       0.211       0.057       0.358       3.39
## 3       17.50        36.6       0.212       0.118       0.395       3.56
## 4       14.30        29.9       0.220       0.097       0.433       3.59
## 5       13.80        36.2       0.165       0.044       0.525       3.37
## 6       13.10        48.4       0.205       0.120       0.346       2.63
##   PC.aa.C40.5 PC.aa.C40.6 PC.aa.C42.0 PC.aa.C42.1 PC.aa.C42.2 PC.aa.C42.4
## 1        5.66        21.8       0.364       0.226       0.108       0.272
## 2        4.08        14.2       0.419       0.216       0.109       0.336
## 3        5.34        16.7       0.476       0.281       0.118       0.300
## 4        5.06        14.0       0.427       0.223       0.119       0.268
## 5        5.29        22.5       0.125       0.095       0.083       0.206
## 6        3.25        18.9       0.451       0.233       0.135       0.228
##   PC.aa.C42.5 PC.aa.C42.6 PC.ae.C30.0 PC.ae.C30.1 PC.ae.C30.2 PC.ae.C32.1
## 1       0.272       0.291       0.173       0.027       0.022        1.65
## 2       0.317       0.248       0.147       0.024       0.020        2.01
## 3       0.206       0.267       0.209       0.046       0.030        2.40
## 4       0.267       0.254       0.223       0.049       0.023        2.47
## 5       0.205       0.280       0.095       0.082       0.023        1.72
## 6       0.254       0.271       0.221       0.039       0.029        2.01
##   PC.ae.C32.2 PC.ae.C34.0 PC.ae.C34.1 PC.ae.C34.2 PC.ae.C34.3 PC.ae.C36.0
## 1       0.371       0.880        3.66        2.48       0.813       0.498
## 2       0.360       0.763        2.68        2.32       0.905       0.398
## 3       0.477       0.938        4.04        2.95       1.030       0.554
## 4       0.459       0.964        4.06        3.09       1.020       0.552
## 5       0.316       1.060        3.28        1.70       0.722       0.553
## 6       0.397       0.920        3.26        2.58       1.000       0.443
##   PC.ae.C36.1 PC.ae.C36.2 PC.ae.C36.3 PC.ae.C36.4 PC.ae.C36.5 PC.ae.C38.0
## 1        5.64        1.90       1.170        6.96        4.79       0.474
## 2        3.89        1.54       0.873        6.40        5.36       0.325
## 3        5.95        2.29       1.240        9.05        6.63       0.478
```

```
## 4        4.75        2.01       1.350        8.36        5.97       0.397
## 5        5.95        1.47       0.760        4.78        4.00       0.430
## 6        4.95        2.05       1.170        7.04        4.47       0.590
##   PC.ae.C38.2 PC.ae.C38.3 PC.ae.C38.4 PC.ae.C38.5 PC.ae.C38.6 PC.ae.C40.1
## 1       0.538        2.66        6.33        5.51        1.95       0.574
## 2       0.127        1.80        5.37        4.49        1.63       0.281
## 3       0.154        2.87        7.06        5.64        1.98       0.759
## 4       0.144        1.97        5.99        5.63        1.97       0.425
## 5       0.246        1.80        5.45        4.34        1.51       0.430
## 6       0.312        2.46        5.55        4.60        1.80       0.481
##   PC.ae.C40.2 PC.ae.C40.3 PC.ae.C40.4 PC.ae.C40.5 PC.ae.C40.6 PC.ae.C42.0
## 1       0.575       0.940        1.76        1.77        1.59       0.629
## 2       0.491       0.702        1.43        1.55        1.20       0.616
## 3       0.654       0.817        1.51        1.64        1.49       0.686
## 4       0.540       0.742        1.45        1.62        1.25       0.637
## 5       0.432       0.632        1.10        1.25        1.47       0.660
## 6       0.598       0.826        1.25        1.38        1.61       0.669
##   PC.ae.C42.1 PC.ae.C42.2 PC.ae.C42.3 PC.ae.C42.4 PC.ae.C42.5 PC.ae.C44.3
## 1       0.316       0.192       0.277       0.264       0.888       0.065
## 2       0.260       0.157       0.200       0.311       0.840       0.071
## 3       0.356       0.241       0.288       0.319       0.957       0.065
## 4       0.299       0.159       0.208       0.392       0.863       0.069
## 5       0.355       0.138       0.174       0.162       0.513       0.081
## 6       0.265       0.195       0.253       0.316       0.814       0.085
##   PC.ae.C44.4 PC.ae.C44.5 PC.ae.C44.6 SM..OH..C14.1 SM..OH..C16.1 SM..OH..C22.1
## 1       0.168       0.536       0.494         1.420          1.33          2.07
## 2       0.220       0.470       0.515         1.390          1.25          2.47
## 3       0.228       0.565       0.603         1.840          1.58          2.69
## 4       0.237       0.517       0.611         1.720          1.48          2.97
## 5       0.154       0.178       0.134         0.987          1.48          1.96
## 6       0.232       0.554       0.539         1.320          1.12          2.51
##   SM..OH..C22.2 SM..OH..C24.1 SM.C16.0 SM.C16.1 SM.C18.0 SM.C18.1 SM.C20.2
## 1          1.86         0.597     44.9     7.99     14.5    10.40    0.290
## 2          2.20         0.640     42.1     6.88     12.7     8.52    0.211
## 3          2.63         0.665     44.8     8.91     14.6    11.60    0.304
## 4          2.84         0.682     52.4     8.61     17.2    11.50    0.261
## 5          1.74         0.478     40.6     5.86     13.0     8.34    0.196
## 6          2.16         0.640     42.6     8.49     13.0    10.60    0.270
##   SM.C24.0 SM.C24.1 SM.C26.0 SM.C26.1 H1_1   H1 Urea_N L.Arginine_N L.Leucine_N
## 1    12.20     27.3    0.147    0.337 3356 3356     NA          NA          NA
## 2    10.40     25.6    0.130    0.317 2509 2509  201.9        22.5        35.3
## 3    11.50     28.8    0.163    0.364 2661 2661  193.3        21.0        25.4
## 4    11.80     27.9    0.138    0.353 2652 2652  500.8        16.0        27.1
## 5     9.29     20.5    0.111    0.283 2258 2258  132.5        13.2        57.9
## 6     9.58     23.7    0.135    0.316 3031 3031  193.3        32.2        26.5
##   EDTAca_N X2.Hydroxybutyrate X3.Hydroxybutyrate Acetate Acetoacetate Acetone
## 1       NA                 NA                 NA      NA           NA      NA
## 2      2.0              12.40                8.5    13.2          5.7     5.1
## 3      1.8              11.33               11.7     5.8          9.3     5.6
## 4      2.5              12.70                7.2     9.8          4.8     4.0
## 5      2.5              35.20               44.7    20.2         18.9    18.9
## 6      0.0              17.20               16.0    23.6          7.8     5.5
##   Betaine Carnitine Choline Creatine Dimethyl.sulfone Ethanol Formate Glucose
## 1      NA        NA      NA       NA               NA      NA      NA      NA
```

```
## 2    22.0       8.7    14.2     14.5             4.7    16.6    24.6  1489.7
## 3    19.1      15.3    14.5     17.8             2.1     8.1    27.4  1343.9
## 4    13.9       7.7    11.8     14.7             1.3     6.4    14.4   629.5
## 5    33.9      18.5    27.7     35.4             5.5    13.0    40.0  1618.0
## 6    16.9      16.7    25.9     18.6             3.4     5.0    35.5  1791.8
##   Glycerol Hypoxanthine Isobutyrate Isopropanol Lactate Malonate
## 1       NA           NA          NA          NA      NA       NA
## 2    324.6          6.3         3.6         1.9  1171.6     10.4
## 3    201.3          6.0         2.5         2.5  1938.1     13.1
## 4    322.0          8.6         2.5         4.4  1037.7      7.6
## 5    271.6          0.0         6.1        11.2  2199.9     11.7
## 6    274.2          8.8         2.3         2.4  1486.7     11.8
```

```
# Replace missing values with the median value of the existing values of that particular column
clean_metabolite2 <- clean_metabolite %>% mutate_all(~ifelse(is.na(.x), median(.x, na.rm = TRUE), .x))
glimpse(clean_metabolite2)
```

```
## Rows: 69
## Columns: 188
## $ Label              <chr> "Alzheimer", "Alzheimer", "Alzheimer", "Alzheimer",~
## $ Phe                <dbl> 72.8, 93.4, 68.6, 94.1, 79.8, 82.5, 69.7, 83.6, 73.~
## $ Pro                <dbl> 166.0, 138.0, 161.0, 129.0, 126.0, 167.0, 95.6, 119~
## $ Ser                <dbl> 170, 142, 158, 162, 115, 173, 143, 135, 145, 174, 1~
## $ Thr                <int> 282, 217, 208, 201, 199, 333, 244, 268, 307, 269, 2~
## $ ADMA               <dbl> 1.150, 1.050, 1.000, 1.100, 1.240, 1.350, 0.991, 1.~
## $ alpha.AAA          <dbl> 0.760, 0.929, 0.620, 0.795, 1.360, 1.150, 0.927, 0.~
## $ c4.OH.Pro          <dbl> 0.236, 0.189, 0.198, 0.198, 0.198, 0.198, 0.184, 0.~
## $ Carnosine          <dbl> 1.270, 1.350, 0.998, 0.675, 1.280, 1.010, 0.702, 0.~
## $ Creatinine         <dbl> 49.9, 48.8, 30.4, 80.1, 60.5, 24.0, 41.6, 30.6, 39.~
## $ DOPA               <dbl> 0.265, 0.252, 0.268, 0.264, 0.271, 0.275, 0.260, 0.~
## $ Dopamine           <dbl> 0.233, 0.231, 0.231, 0.234, 0.231, 0.231, 0.231, 0.~
## $ Histamine          <dbl> 0.225, 0.211, 0.217, 0.209, 0.210, 0.212, 0.211, 0.~
## $ Kynurenine         <dbl> 5.21, 5.44, 5.20, 5.80, 4.46, 7.01, 6.18, 5.66, 6.3~
## $ Met.SO             <dbl> 0.526, 0.387, 0.651, 0.389, 0.466, 0.417, 0.358, 0.~
## $ Putrescine         <dbl> 0.068, 0.087, 0.260, 0.110, 0.118, 0.262, 0.176, 0.~
## $ Sarcosine          <dbl> 17.8, 20.2, 14.4, 18.7, 22.5, 30.8, 16.3, 23.3, 22.~
## $ Serotonin          <dbl> 0.147, 0.231, 0.196, 0.255, 0.390, 0.140, 0.162, 0.~
## $ Spermidine         <dbl> 0.188, 0.233, 0.384, 0.353, 0.473, 0.856, 0.060, 0.~
## $ t4.OH.Pro          <dbl> 24.0, 29.3, 20.9, 23.1, 26.9, 26.0, 15.7, 10.7, 16.~
## $ Taurine            <dbl> 125, 120, 139, 159, 149, 379, 168, 133, 215, 140, 3~
## $ SDMA               <dbl> 1.13, 1.65, 1.57, 1.34, 1.24, 1.44, 1.32, 1.04, 1.2~
## $ C0                 <dbl> 18.2, 17.0, 12.6, 23.5, 13.6, 26.7, 12.9, 13.3, 15.~
## $ C10                <dbl> 0.059, 0.051, 0.083, 0.071, 0.139, 0.058, 0.063, 0.~
## $ C10.1              <dbl> 0.312, 0.288, 0.357, 0.317, 0.472, 0.238, 0.247, 0.~
## $ C10.2              <dbl> 0.038, 0.039, 0.054, 0.040, 0.074, 0.042, 0.041, 0.~
## $ C12                <dbl> 0.030, 0.038, 0.032, 0.045, 0.056, 0.039, 0.037, 0.~
## $ C12.DC             <dbl> 0.042, 0.038, 0.048, 0.048, 0.079, 0.035, 0.038, 0.~
## $ C12.1              <dbl> 0.290, 0.265, 0.302, 0.275, 0.394, 0.196, 0.218, 0.~
## $ C14                <dbl> 0.023, 0.026, 0.021, 0.026, 0.034, 0.029, 0.025, 0.~
## $ C14.1              <dbl> 0.019, 0.017, 0.031, 0.028, 0.043, 0.023, 0.029, 0.~
## $ C14.1.OH           <dbl> 0.008, 0.008, 0.010, 0.010, 0.016, 0.009, 0.008, 0.~
## $ C14.2              <dbl> 0.008, 0.009, 0.010, 0.013, 0.025, 0.010, 0.011, 0.~
## $ C14.2.OH           <dbl> 0.006, 0.009, 0.009, 0.011, 0.017, 0.007, 0.008, 0.~
## $ C16                <dbl> 0.046, 0.070, 0.076, 0.074, 0.062, 0.081, 0.057, 0.~
```

```
## $ C16.OH             <dbl> 0.008, 0.009, 0.011, 0.011, 0.007, 0.006, 0.007, 0.~
## $ C16.1              <dbl> 0.009, 0.013, 0.019, 0.015, 0.024, 0.012, 0.013, 0.~
## $ C16.1.OH           <dbl> 0.007, 0.006, 0.010, 0.008, 0.014, 0.005, 0.007, 0.~
## $ C16.2              <dbl> 0.005, 0.006, 0.005, 0.006, 0.012, 0.007, 0.005, 0.~
## $ C16.2.OH           <dbl> 0.013, 0.012, 0.013, 0.009, 0.025, 0.015, 0.011, 0.~
## $ C18                <dbl> 0.013, 0.014, 0.016, 0.020, 0.031, 0.017, 0.019, 0.~
## $ C18.1              <dbl> 0.024, 0.025, 0.025, 0.035, 0.034, 0.035, 0.037, 0.~
## $ C18.1.OH           <dbl> 0.003, 0.003, 0.004, 0.004, 0.012, 0.004, 0.004, 0.~
## $ C18.2              <dbl> 0.016, 0.028, 0.018, 0.033, 0.017, 0.029, 0.018, 0.~
## $ C2                 <dbl> 1.97, 1.95, 1.70, 2.10, 5.62, 3.49, 2.17, 1.66, 2.2~
## $ C3                 <dbl> 0.354, 0.184, 0.371, 0.278, 0.436, 0.461, 0.253, 0.~
## $ C3.OH              <dbl> 0.008, 0.009, 0.011, 0.010, 0.029, 0.008, 0.009, 0.~
## $ C3.1               <dbl> 0.015, 0.013, 0.012, 0.017, 0.035, 0.014, 0.015, 0.~
## $ C4                 <dbl> 0.082, 0.108, 0.057, 0.110, 0.106, 0.123, 0.068, 0.~
## $ C3.DC..C4.OH.      <dbl> 0.045, 0.080, 0.035, 0.077, 0.099, 0.068, 0.066, 0.~
## $ C4.1               <dbl> 0.025, 0.025, 0.039, 0.031, 0.069, 0.026, 0.014, 0.~
## $ C5                 <dbl> 0.094, 0.077, 0.096, 0.145, 0.141, 0.090, 0.077, 0.~
## $ C5.M.DC            <dbl> 0.023, 0.032, 0.045, 0.034, 0.094, 0.019, 0.030, 0.~
## $ C5.OH..C3.DC.M.    <dbl> 0.026, 0.026, 0.024, 0.041, 0.058, 0.037, 0.022, 0.~
## $ C5.1               <dbl> 0.030, 0.024, 0.037, 0.035, 0.073, 0.022, 0.020, 0.~
## $ C5.1.DC            <dbl> 0.020, 0.021, 0.018, 0.016, 0.049, 0.016, 0.016, 0.~
## $ C6..C4.1.DC.       <dbl> 0.022, 0.030, 0.022, 0.029, 0.052, 0.063, 0.029, 0.~
## $ C5.DC..C6.OH.      <dbl> 0.014, 0.018, 0.029, 0.016, 0.040, 0.016, 0.016, 0.~
## $ C6.1               <dbl> 0.018, 0.015, 0.031, 0.027, 0.040, 0.019, 0.017, 0.~
## $ C7.DC              <dbl> 0.011, 0.010, 0.021, 0.017, 0.036, 0.014, 0.014, 0.~
## $ C8                 <dbl> 0.062, 0.058, 0.090, 0.091, 0.192, 0.073, 0.056, 0.~
## $ C9                 <dbl> 0.016, 0.014, 0.017, 0.018, 0.041, 0.014, 0.014, 0.~
## $ lysoPC.a.C14.0     <dbl> 2.23, 1.97, 2.12, 2.19, 1.88, 2.11, 2.32, 2.13, 2.1~
## $ lysoPC.a.C16.0     <dbl> 37.9, 22.1, 33.7, 32.8, 24.5, 29.1, 42.4, 33.7, 36.~
## $ lysoPC.a.C16.1     <dbl> 2.66, 1.31, 2.53, 2.39, 1.27, 2.09, 3.16, 3.09, 3.4~
## $ lysoPC.a.C17.0     <dbl> 0.446, 0.270, 0.399, 0.323, 0.382, 0.348, 0.437, 0.~
## $ lysoPC.a.C18.0     <dbl> 9.00, 5.35, 7.51, 7.21, 6.66, 5.84, 9.63, 6.96, 7.2~
## $ lysoPC.a.C18.1     <dbl> 8.58, 3.94, 7.73, 7.22, 5.39, 6.30, 9.44, 7.31, 8.1~
## $ lysoPC.a.C18.2     <dbl> 7.27, 4.42, 8.02, 7.62, 3.60, 8.10, 10.90, 7.53, 6.~
## $ lysoPC.a.C20.3     <dbl> 1.830, 0.958, 2.050, 1.640, 0.970, 1.970, 2.540, 2.~
## $ lysoPC.a.C20.4     <dbl> 8.25, 4.60, 9.84, 6.75, 6.26, 7.04, 10.80, 8.73, 7.~
## $ lysoPC.a.C24.0     <dbl> 0.079, 0.059, 0.075, 0.066, 0.084, 0.083, 0.069, 0.~
## $ lysoPC.a.C26.0     <dbl> 0.113, 0.066, 0.126, 0.086, 0.118, 0.112, 0.095, 0.~
## $ lysoPC.a.C26.1     <dbl> 0.053, 0.042, 0.049, 0.045, 0.053, 0.050, 0.049, 0.~
## $ lysoPC.a.C28.0     <dbl> 0.108, 0.076, 0.078, 0.076, 0.092, 0.099, 0.107, 0.~
## $ lysoPC.a.C28.1     <dbl> 0.072, 0.058, 0.092, 0.076, 0.072, 0.083, 0.088, 0.~
## $ PC.aa.C24.0        <dbl> 0.082, 0.065, 0.099, 0.076, 0.069, 0.073, 0.074, 0.~
## $ PC.aa.C26.0        <dbl> 0.438, 0.409, 0.458, 0.486, 0.401, 0.450, 0.424, 0.~
## $ PC.aa.C28.1        <dbl> 0.571, 0.521, 0.605, 0.685, 0.513, 0.620, 0.788, 0.~
## $ PC.aa.C30.0        <dbl> 2.35, 1.99, 2.69, 3.33, 1.78, 2.61, 2.42, 2.32, 2.0~
## $ PC.aa.C32.0        <dbl> 11.40, 12.70, 16.60, 18.60, 13.80, 14.70, 12.40, 12~
## $ PC.aa.C32.1        <dbl> 9.22, 5.40, 11.60, 13.30, 5.03, 8.98, 10.40, 11.50,~
## $ PC.aa.C32.2        <dbl> 0.117, 0.117, 0.117, 0.053, 0.117, 0.117, 0.117, 0.~
## $ PC.aa.C32.3        <dbl> 0.092, 0.067, 0.105, 0.079, 0.102, 0.107, 0.121, 0.~
## $ PC.aa.C34.1        <dbl> 109.0, 64.2, 108.0, 106.0, 83.4, 90.2, 111.0, 83.6,~
## $ PC.aa.C34.2        <dbl> 71.0, 60.5, 83.1, 93.6, 35.9, 85.6, 92.7, 60.6, 55.~
## $ PC.aa.C34.3        <dbl> 1.430, 0.879, 1.930, 1.590, 0.709, 1.790, 2.040, 1.~
## $ PC.aa.C34.4        <dbl> 0.200, 0.127, 0.210, 0.190, 0.135, 0.213, 0.315, 0.~
## $ PC.aa.C36.0        <dbl> 2.38, 2.05, 2.30, 2.57, 1.83, 2.48, 2.22, 2.16, 1.6~
```

```
## $ PC.aa.C36.1       <dbl> 21.7, 14.3, 19.9, 20.9, 20.5, 15.5, 21.3, 18.4, 18.~
## $ PC.aa.C36.2       <dbl> 42.4, 35.6, 44.9, 48.8, 28.5, 43.2, 55.3, 34.4, 32.~
## $ PC.aa.C36.3       <dbl> 42.7, 24.3, 43.9, 41.2, 21.9, 46.0, 54.9, 41.5, 41.~
## $ PC.aa.C36.4       <dbl> 120.0, 83.7, 146.0, 122.0, 98.1, 114.0, 137.0, 110.~
## $ PC.aa.C36.5       <dbl> 1.86, 1.05, 2.09, 1.76, 1.70, 3.47, 2.46, 2.03, 1.7~
## $ PC.aa.C36.6       <dbl> 0.084, 0.046, 0.057, 0.070, 0.048, 0.103, 0.113, 0.~
## $ PC.aa.C38.0       <dbl> 1.230, 0.946, 1.210, 1.160, 1.100, 1.390, 1.110, 1.~
## $ PC.aa.C38.3       <dbl> 32.1, 21.9, 34.5, 28.7, 23.3, 28.9, 42.4, 31.3, 31.~
## $ PC.aa.C38.4       <dbl> 95.1, 78.9, 107.0, 92.7, 101.0, 78.0, 109.0, 81.7, ~
## $ PC.aa.C38.5       <dbl> 16.80, 9.91, 17.50, 14.30, 13.80, 13.10, 17.60, 14.~
## $ PC.aa.C38.6       <dbl> 41.6, 25.1, 36.6, 29.9, 36.2, 48.4, 46.0, 42.8, 37.~
## $ PC.aa.C40.1       <dbl> 0.195, 0.211, 0.212, 0.220, 0.165, 0.205, 0.192, 0.~
## $ PC.aa.C40.2       <dbl> 0.074, 0.057, 0.118, 0.097, 0.044, 0.120, 0.039, 0.~
## $ PC.aa.C40.3       <dbl> 0.491, 0.358, 0.395, 0.433, 0.525, 0.346, 0.392, 0.~
## $ PC.aa.C40.4       <dbl> 3.48, 3.39, 3.56, 3.59, 3.37, 2.63, 3.52, 4.02, 2.8~
## $ PC.aa.C40.5       <dbl> 5.66, 4.08, 5.34, 5.06, 5.29, 3.25, 5.79, 5.49, 4.8~
## $ PC.aa.C40.6       <dbl> 21.80, 14.20, 16.70, 14.00, 22.50, 18.90, 22.70, 20~
## $ PC.aa.C42.0       <dbl> 0.364, 0.419, 0.476, 0.427, 0.125, 0.451, 0.468, 0.~
## $ PC.aa.C42.1       <dbl> 0.226, 0.216, 0.281, 0.223, 0.095, 0.233, 0.247, 0.~
## $ PC.aa.C42.2       <dbl> 0.108, 0.109, 0.118, 0.119, 0.083, 0.135, 0.119, 0.~
## $ PC.aa.C42.4       <dbl> 0.272, 0.336, 0.300, 0.268, 0.206, 0.228, 0.225, 0.~
## $ PC.aa.C42.5       <dbl> 0.272, 0.317, 0.206, 0.267, 0.205, 0.254, 0.226, 0.~
## $ PC.aa.C42.6       <dbl> 0.291, 0.248, 0.267, 0.254, 0.280, 0.271, 0.297, 0.~
## $ PC.ae.C30.0       <dbl> 0.173, 0.147, 0.209, 0.223, 0.095, 0.221, 0.191, 0.~
## $ PC.ae.C30.1       <dbl> 0.027, 0.024, 0.046, 0.049, 0.082, 0.039, 0.012, 0.~
## $ PC.ae.C30.2       <dbl> 0.022, 0.020, 0.030, 0.023, 0.023, 0.029, 0.032, 0.~
## $ PC.ae.C32.1       <dbl> 1.65, 2.01, 2.40, 2.47, 1.72, 2.01, 1.70, 1.68, 1.5~
## $ PC.ae.C32.2       <dbl> 0.371, 0.360, 0.477, 0.459, 0.316, 0.397, 0.369, 0.~
## $ PC.ae.C34.0       <dbl> 0.880, 0.763, 0.938, 0.964, 1.060, 0.920, 0.723, 1.~
## $ PC.ae.C34.1       <dbl> 3.66, 2.68, 4.04, 4.06, 3.28, 3.26, 3.69, 3.51, 3.2~
## $ PC.ae.C34.2       <dbl> 2.48, 2.32, 2.95, 3.09, 1.70, 2.58, 2.46, 2.28, 2.0~
## $ PC.ae.C34.3       <dbl> 0.813, 0.905, 1.030, 1.020, 0.722, 1.000, 0.881, 0.~
## $ PC.ae.C36.0       <dbl> 0.498, 0.398, 0.554, 0.552, 0.553, 0.443, 0.457, 0.~
## $ PC.ae.C36.1       <dbl> 5.64, 3.89, 5.95, 4.75, 5.95, 4.95, 5.59, 5.65, 4.7~
## $ PC.ae.C36.2       <dbl> 1.90, 1.54, 2.29, 2.01, 1.47, 2.05, 2.25, 1.97, 1.5~
## $ PC.ae.C36.3       <dbl> 1.170, 0.873, 1.240, 1.350, 0.760, 1.170, 1.370, 1.~
## $ PC.ae.C36.4       <dbl> 6.96, 6.40, 9.05, 8.36, 4.78, 7.04, 7.56, 7.15, 6.4~
## $ PC.ae.C36.5       <dbl> 4.79, 5.36, 6.63, 5.97, 4.00, 4.47, 4.69, 4.04, 3.3~
## $ PC.ae.C38.0       <dbl> 0.474, 0.325, 0.478, 0.397, 0.430, 0.590, 0.583, 0.~
## $ PC.ae.C38.2       <dbl> 0.538, 0.127, 0.154, 0.144, 0.246, 0.312, 0.065, 0.~
## $ PC.ae.C38.3       <dbl> 2.66, 1.80, 2.87, 1.97, 1.80, 2.46, 2.81, 2.90, 2.5~
## $ PC.ae.C38.4       <dbl> 6.33, 5.37, 7.06, 5.99, 5.45, 5.55, 6.03, 5.73, 5.0~
## $ PC.ae.C38.5       <dbl> 5.51, 4.49, 5.64, 5.63, 4.34, 4.60, 4.88, 4.53, 3.8~
## $ PC.ae.C38.6       <dbl> 1.95, 1.63, 1.98, 1.97, 1.51, 1.80, 1.72, 1.71, 1.2~
## $ PC.ae.C40.1       <dbl> 0.574, 0.281, 0.759, 0.425, 0.430, 0.481, 0.744, 0.~
## $ PC.ae.C40.2       <dbl> 0.575, 0.491, 0.654, 0.540, 0.432, 0.598, 0.803, 0.~
## $ PC.ae.C40.3       <dbl> 0.940, 0.702, 0.817, 0.742, 0.632, 0.826, 0.871, 0.~
## $ PC.ae.C40.4       <dbl> 1.76, 1.43, 1.51, 1.45, 1.10, 1.25, 1.28, 1.84, 1.3~
## $ PC.ae.C40.5       <dbl> 1.77, 1.55, 1.64, 1.62, 1.25, 1.38, 1.51, 1.53, 1.3~
## $ PC.ae.C40.6       <dbl> 1.590, 1.200, 1.490, 1.250, 1.470, 1.610, 1.440, 1.~
## $ PC.ae.C42.0       <dbl> 0.629, 0.616, 0.686, 0.637, 0.660, 0.669, 0.679, 0.~
## $ PC.ae.C42.1       <dbl> 0.316, 0.260, 0.356, 0.299, 0.355, 0.265, 0.350, 0.~
## $ PC.ae.C42.2       <dbl> 0.192, 0.157, 0.241, 0.159, 0.138, 0.195, 0.215, 0.~
## $ PC.ae.C42.3       <dbl> 0.277, 0.200, 0.288, 0.208, 0.174, 0.253, 0.271, 0.~
```

```
## $ PC.ae.C42.4        <dbl> 0.264, 0.311, 0.319, 0.392, 0.162, 0.316, 0.316, 0.~
## $ PC.ae.C42.5        <dbl> 0.888, 0.840, 0.957, 0.863, 0.513, 0.814, 0.936, 0.~
## $ PC.ae.C44.3        <dbl> 0.065, 0.071, 0.065, 0.069, 0.081, 0.085, 0.069, 0.~
## $ PC.ae.C44.4        <dbl> 0.168, 0.220, 0.228, 0.237, 0.154, 0.232, 0.199, 0.~
## $ PC.ae.C44.5        <dbl> 0.536, 0.470, 0.565, 0.517, 0.178, 0.554, 0.598, 0.~
## $ PC.ae.C44.6        <dbl> 0.494, 0.515, 0.603, 0.611, 0.134, 0.539, 0.542, 0.~
## $ SM..OH..C14.1      <dbl> 1.420, 1.390, 1.840, 1.720, 0.987, 1.320, 1.900, 1.~
## $ SM..OH..C16.1      <dbl> 1.330, 1.250, 1.580, 1.480, 1.480, 1.120, 1.640, 1.~
## $ SM..OH..C22.1      <dbl> 2.07, 2.47, 2.69, 2.97, 1.96, 2.51, 3.00, 2.98, 2.2~
## $ SM..OH..C22.2      <dbl> 1.86, 2.20, 2.63, 2.84, 1.74, 2.16, 2.89, 2.59, 2.0~
## $ SM..OH..C24.1      <dbl> 0.597, 0.640, 0.665, 0.682, 0.478, 0.640, 0.690, 0.~
## $ SM.C16.0           <dbl> 44.9, 42.1, 44.8, 52.4, 40.6, 42.6, 47.2, 37.9, 37.~
## $ SM.C16.1           <dbl> 7.99, 6.88, 8.91, 8.61, 5.86, 8.49, 8.63, 7.92, 6.6~
## $ SM.C18.0           <dbl> 14.5, 12.7, 14.6, 17.2, 13.0, 13.0, 18.6, 11.9, 12.~
## $ SM.C18.1           <dbl> 10.40, 8.52, 11.60, 11.50, 8.34, 10.60, 13.10, 9.59~
## $ SM.C20.2           <dbl> 0.290, 0.211, 0.304, 0.261, 0.196, 0.270, 0.349, 0.~
## $ SM.C24.0           <dbl> 12.20, 10.40, 11.50, 11.80, 9.29, 9.58, 11.40, 9.36~
## $ SM.C24.1           <dbl> 27.3, 25.6, 28.8, 27.9, 20.5, 23.7, 28.5, 18.8, 23.~
## $ SM.C26.0           <dbl> 0.147, 0.130, 0.163, 0.138, 0.111, 0.135, 0.140, 0.~
## $ SM.C26.1           <dbl> 0.337, 0.317, 0.364, 0.353, 0.283, 0.316, 0.386, 0.~
## $ H1_1               <int> 3356, 2509, 2661, 2652, 2258, 3031, 2688, 2464, 272~
## $ H1                 <int> 3356, 2509, 2661, 2652, 2258, 3031, 2688, 2464, 272~
## $ Urea_N             <dbl> 185.05, 201.90, 193.30, 500.80, 132.50, 193.30, 159~
## $ L.Arginine_N       <dbl> 45.1, 22.5, 21.0, 16.0, 13.2, 32.2, 59.6, 49.8, 39.~
## $ L.Leucine_N        <dbl> 55.75, 35.30, 25.40, 27.10, 57.90, 26.50, 61.20, 63~
## $ EDTAca_N           <dbl> 2.9, 2.0, 1.8, 2.5, 2.5, 0.0, 2.3, 0.0, 2.7, 2.3, 0~
## $ X2.Hydroxybutyrate <dbl> 19.80, 12.40, 11.33, 12.70, 35.20, 17.20, 45.60, 21~
## $ X3.Hydroxybutyrate <dbl> 44.10, 8.50, 11.70, 7.20, 44.70, 16.00, 22.65, 20.9~
## $ Acetate            <dbl> 20.2, 13.2, 5.8, 9.8, 20.2, 23.6, 22.3, 19.5, 20.0,~
## $ Acetoacetate       <dbl> 21.4, 5.7, 9.3, 4.8, 18.9, 7.8, 91.0, 15.4, 22.0, 2~
## $ Acetone            <dbl> 10.15, 5.10, 5.60, 4.00, 18.90, 5.50, 28.40, 6.60, ~
## $ Betaine            <dbl> 32.25, 22.00, 19.10, 13.90, 33.90, 16.90, 37.50, 35~
## $ Carnitine          <dbl> 13.1, 8.7, 15.3, 7.7, 18.5, 16.7, 4.8, 13.0, 14.4, ~
## $ Choline            <dbl> 22.15, 14.20, 14.50, 11.80, 27.70, 25.90, 20.10, 21~
## $ Creatine           <dbl> 26.7, 14.5, 17.8, 14.7, 35.4, 18.6, 25.4, 25.9, 25.~
## $ Dimethyl.sulfone   <dbl> 3.55, 4.70, 2.10, 1.30, 5.50, 3.40, 3.70, 5.40, 3.5~
## $ Ethanol            <dbl> 7.2, 16.6, 8.1, 6.4, 13.0, 5.0, 6.3, 10.2, 5.1, 4.4~
## $ Formate            <dbl> 28.9, 24.6, 27.4, 14.4, 40.0, 35.5, 27.6, 23.2, 25.~
## $ Glucose            <dbl> 2239.35, 1489.70, 1343.90, 629.50, 1618.00, 1791.80~
## $ Glycerol           <dbl> 449.1, 324.6, 201.3, 322.0, 271.6, 274.2, 619.7, 40~
## $ Hypoxanthine       <dbl> 7.35, 6.30, 6.00, 8.60, 0.00, 8.80, 6.90, 5.80, 5.6~
## $ Isobutyrate        <dbl> 4.6, 3.6, 2.5, 2.5, 6.1, 2.3, 5.0, 4.5, 5.9, 5.5, 4~
## $ Isopropanol        <dbl> 3.3, 1.9, 2.5, 4.4, 11.2, 2.4, 1.8, 4.4, 6.7, 2.7, ~
## $ Lactate            <dbl> 1768.7, 1171.6, 1938.1, 1037.7, 2199.9, 1486.7, 204~
## $ Malonate           <dbl> 11.35, 10.40, 13.10, 7.60, 11.70, 11.80, 9.70, 11.0~
```

6: Grad Students Only Please apply Principal Component Analysis (PCA) on the processed metabolites data and create a scatter plot by using first two principal components in which points are colored based on the Label column. Please submit your code along with your figure in the same file.

(If you are going to use R, you may need to use which(), is.na() functions and consider excluding those columns by name. For that purpose you may investigate %in% and -c(...) type of operations. You can also see examples of subsetting a dataframe below with their outputs. It's also recommended to check tidyverse library.)

```r
# Using R
# Apply PCA on the processed metabolites data
pca_metabolite <- clean_metabolite2 %>% select(Phe ,Pro,) %>% prcomp(scale = TRUE)
print(pca_metabolite)
```

```
## Standard deviations (1, .., p=2):
## [1] 1.1391092 0.8381111
##
## Rotation (n x k) = (2 x 2):
##           PC1        PC2
## Phe 0.7071068  0.7071068
## Pro 0.7071068 -0.7071068
```

```r
# Create a scatter plot by using first two principal components
pca_metabolite_df <- as.data.frame(pca_metabolite$x)
pca_metabolite_df$Label <- clean_metabolite2$Label
ggplot(data = pca_metabolite_df, aes(x = PC1, y = PC2, color = Label)) + geom_point()
```