DATA MINING WITH R

A S SAI THEJASWINI

DATA MINING

DATA MINING IS THE COMPUTING PROCESS OF DISCOVERING PATTERNS IN LARGE DATASETS INVOLVING THE METHODS AT THE INTERSECTION OF MACHINE LEARNING, STATISTICS, AND DATABASE SYSTEMS.

DATA MINING TOOLS ARE USED TO EXTRACT KNOWLEDGE FROM COLLABRATION PATEERNS FROM SYSTEM LOGS AND ANALYZE THE IMPACT OF COLLABRATION PATTERNS ON PROCESS EFFICIENCY.

OBJECTIVE OF THE PROJECT

THE OBJECTIVE OF THE PROJECT IS TO ANALYZE THE DATASET AND EXTRACT USEFUL INFORMATION, PERFORM ANALYSIS FOR FINDING THE MISSING VALUES, FILLING THE MISSING VALUES AND OUTLIER DETECTION, IDENTIFYING THE CLUSTERS AND THEE PATTERNS IN THE DATASET.

DATASET

THE DATASET CHOOSEN FOR THE PROJECT IS A HEALTH CARE DATASET, BASED ON THE HEART ATTACK PREDICTION.

HEART ATTACK PREDICTION IS BASED ON THE AGE , BMI , OBESITY, CHOLESTROL , BLOOD PRESSURE, DIABETICS, ETC. ON THE DATASET.

STEPS PERFORMED IN THE DATASET.

- BASIC STATISTICS : THE DESCRIPTIVE STATISTICS OF THE MEAN, MEDIAN VARIANCE,SD, QURATILE , IQR AND SUMMARY OF THE DATASET IS PERFORMED.
- FINDING THE MISSING VALUES : FINDING THE MISSING VALUES IN THE DATASET(IF ANY) IS PERFORMED AS PART OF THE PRE-PROCESSING.
- FILLING THE NA VALUES : THE NA VALUES IS FILLED WITH THE MEAN VALUE OF THE COLUMN AS PART OF THE PRE-PROCESSING.

- LABEL ENCODING : THE DIET COLUMN IN THE DATASET IS ENCODED WITH THE LABEL VAUES(0,1,2,ETC) FOR THE ANALYSIS

- ONE HOT ENCODING : THE SEX COLUMN IN THE DATASET IS HAS BEEN PERFORMED WITH THE ONE- HOT ENCODING TO CONVERT THE VALUES TO ZEROS AND ONES.
- DATA BINNING : BINNING IS PERFORMED TO KNOW THE COUNT OF PEOPLE IN BETWEEN THE AGE GROUP.

- FEATURE SCALING : NORMALISING THE DATSEET COLUMNS WITH HIGH RANGE TO THE ZEROS AND ONES.

- FEATURE SELECTION : THE MOST REQUIRED FEATURES FOR THE EXTRACTING PATTERNS ARE SELECTED FROM THE LARGE DATASET
- CORRELATION : THE CORRELATION BETWEEN THE ITEMS PROVIDE THE DEPENDENCY OF THE VARIABLE TO THAT OTHER.

- OUTLIER DETECTION : THE OUTLIERS IN THE DATASET IS BEEN DETECTED AND RECTIFIED FOR THE ANALYSIS USING THE VISUALIZATION .

- HIERARCHICAL CLUSTERING : THE HIERARCHICAL CLUSTERING IS DONE TO THE PATTERNS IN THE DATASET.

DATA MINING WITH R

```r
# IMPORTING THE DATASET.
heart = read.csv(file.choose())
heart

```

	Patient.ID	Age	Sex	Cholesterol	Blood.Pressure	Heart.Rate	Diabetes
1	BMW7812	67	Male	208	158/88	72	0
2	CZE1114	21	Male	389	165/93	98	1
3	BNI9906	21	Female	324	174/99	72	1
4	JLN3497	84	Male	383	163/100	73	1
5	GFO8847	66	Male	318	91/88	93	1
6	ZOO7941	54	Female	297	172/86	48	1
7	WYV0966	90	Male	358	102/73	84	0
8	XXM0972	84	Male	220	131/68	107	0

	Family.History	Smoking	Obesity	Alcohol.Consumption	Exercise.Hours.Per.Week
1	0	1	0	0	4.1681888
2	1	1	1	1	1.8132416
3	0	0	0	0	2.0783530
4	1	1	0	1	9.8281296
5	1	1	1	0	5.8042988
6	1	1	0	1	0.6250080
7	0	1	0	1	4.0981771
8	0	1	1	1	3.4279288
9	0	1	1	0	16.8683022
10	1	1	1	1	0.1945151
11	1	1	0	1	16.8419876
12	1	1	1	1	8.2519951
13	1	1	1	1	19.6332682
14	1	1	0	1	17.0373742
15	1	1	0	1	15.3876046

	Diet	Previous.Heart.Problems	Medication.Use	Stress.Level
1	Average	0	0	9
2	Unhealthy	1	0	1
3	Healthy	1	1	9
4	Average	1	0	9
5	Unhealthy	1	0	6
6	Unhealthy	1	1	2
7	Healthy	0	0	7
8	Average	0	1	4
9	Average	0	0	5
10	Unhealthy	0	0	4
11	Average	1	1	8
12	Average	0	0	4
13	Unhealthy	0	0	9
14	Healthy	1	1	1
15	Unhealthy	0	1	2

	Sedentary.Hours.Per.Day	Income	BMI	Triglycerides
1	6.615001	261404	31.25123	286
2	4.963459	285768	27.19497	235
3	9.463426	235282	28.17657	587
4	7.648981	125640	36.46470	378
5	1.514821	160555	21.80914	231
6	7.798752	241339	20.14684	795
7	0.627356	190450	28.88581	284
8	10.543780	122093	22.22186	370
9	11.348787	25086	35.80990	790
10	4.055115	209703	22.55892	232
11	8.919879	50030	22.86791	469
12	7.227338	163066	32.48535	523
13	10.917524	29886	35.10224	590
14	8.727417	292173	25.56490	506
15	10.425490	165300	25.49174	635

```r
#SOME BASSIC STATISTICS IN R
mean(heart$Age)
median(heart$Age)
var(heart$Cholesterol)
sd(heart$Cholesterol)
range(heart$Cholesterol)
quantile(heart$Age)
IQR(heart$Age)
str(heart)
fivenum(heart$Age)
```

```
> mean(heart$Age)
[1] 53.70433
> median(heart$Age)
[1] 54
> var(heart$Cholesterol)
[1] 6538.869
> sd(heart$Cholesterol)
[1] 80.86328
> range(heart$Cholesterol)
> range(heart$Cholesterol)
[1] 120 400
> quantile(heart$Age)
  0%  25%  50%  75% 100%
  18   35   54   72   90
> IQR(heart$Age)
[1] 37
 $ Sex                             : chr  "Male" "Male" "Female" "Male" ...
 $ Cholesterol                     : int  208 389 324 383 318 297 358 220 145
248 ...
 $ Blood.Pressure                  : chr  "158/88" "165/93" "174/99" "163/100"
...
 $ Heart.Rate                      : int  72 98 72 73 93 48 84 107 68 55 ...
 $ Diabetes                        : int  0 1 1 1 1 1 0 0 1 0 ...
 $ Family.History                  : int  0 1 0 1 1 1 0 0 0 1 ...
 $ Smoking                         : int  1 1 0 1 1 1 1 1 1 1 ...
 $ Obesity                         : int  0 1 0 0 1 0 0 1 1 1 ...
```

```
16  # Finding the missing values if any in the dataset
17  is.na(heart)
18  sum(is.na(heart))
19
```

```
 [33,]   FALSE     FALSE       FALSE                FALSE
 [34,]   FALSE     FALSE       FALSE                FALSE
 [35,]   FALSE     FALSE       FALSE                FALSE
 [36,]   FALSE     FALSE       FALSE                FALSE
 [37,]   FALSE     FALSE       FALSE                FALSE
 [38,]   FALSE     FALSE       FALSE                FALSE
 [ reached getOption("max.print") -- omitted 8725 rows ]
> sum(is.na(heart))
[1] 1
>
```

```
20  #  Replacing the NA value with the  mean  value
21  heart$Age[is.na(heart$Age)]= mean(heart$Age,na.rm=TRUE)
22  sum(is.na(heart))
23
```

```
> #  Replacing the NA value with the  mean  value
> heart$Age[is.na(heart$Age)]= mean(heart$Age,na.rm=TRUE)
> sum(is.na(heart))
[1] 0
>
```

```r
#Label encoding the Diet column
heart$Diet = as.integer(factor(heart$Diet))
heart$Diet
```

```
[579] 3 3 3 1 2 3 1 3 1 1 1 2 2 2 2 2 3 2 2 1 3 1 1 3 2 2 1 3 1 1 2 3 2 3
[613] 2 3 3 1 2 1 1 3 1 2 2 3 1 1 3 1 1 2 1 3 1 2 1 2 3 1 3 1 3 3 3 2 2 1
[647] 3 2 1 2 2 3 1 3 2 3 1 3 3 1 3 3 2 1 1 1 3 2 1 3 1 1 2 3 3 1 2 2 3 2
[681] 3 2 3 3 2 3 2 2 2 2 3 1 1 2 2 2 2 3 2 3 3 3 2 3 2 3 2 1 1 3 1 1 3 1
[715] 2 3 1 2 2 1 1 3 1 1 2 2 2 1 3 1 3 3 2 3 1 2 1 1 2 3 1 2 3 2 3 2 3 2
[749] 3 3 2 3 3 1 2 3 2 3 1 3 2 3 2 3 2 1 2 1 2 2 2 1 2 2 2 2 3 1 3 1 1 3
[783] 1 3 1 1 3 2 1 1 3 2 3 3 1 3 3 2 1 1 2 3 2 1 1 1 2 1 2 3 3 1 3 1 1 3
[817] 3 1 3 3 3 3 2 3 3 3 1 3 3 1 3 1 2 1 2 3 2 3 2 3 1 2 1 3 2 3 1 1 1 3
[851] 2 2 1 1 3 2 3 3 1 3 3 2 2 1 2 3 1 2 1 2 3 3 3 1 1 2 2 3 2 2 2 2 1 3
[885] 2 1 3 3 2 3 3 3 1 2 2 2 2 2 2 2 1 1 3 3 3 1 1 2 3 2 2 3 1 1 2 1 1 1
```

```r
# one - hot encoding in the sex column
heart$Sex <- factor(heart$Sex,
                    levels=c("Female","Male"),
                    labels=c(0,1))
heart$Sex
```

```
> heart$Sex
  [1] 1 1 0 1 1 0 1 1 1 0 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 0 1 1 1
 [35] 1 0 0 0 1 1 0 1 0 1 1 1 1 0 1 0 1 1 1 1 1 1 0 0 1 1 1 1 0 1 1 1 1 0
 [69] 0 1 0 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 0 0 1 1 1 0 0 0 0 1 1 0 1 1 1 1
[103] 1 1 1 1 0 1 0 1 0 1 0 0 1 1 1 0 1 0 1 0 1 1 1 0 1 1 1 0 1 1 0 0 0 1
[137] 1 0 0 1 1 0 0 1 0 1 1 0 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 0 1 1 0 1 0 0
[171] 0 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 0 1 1 1 0 1 1 0 1 0 1 0 0 1 1 0 1
[205] 1 0 0 1 0 1 1 1 1 1 1 0 0 1 1 1 0 1 1 1 0 0 1 1 0 1 1 0 1 1 1 1 1 0 1 1
```

```
34  #  Binning
35  bin_edges = c(0,20,40,60,80)
36
37  data_bins = cut(heart$Age,breaks = bin_edges,include.lowest = TRUE,
38                  labels =  c("Bin1","Bin2","Bin3","Bin4"))
39
40  bin_counts  = table(data_bins)
41
42  print(bin_counts)
```

```
> #  Binning
> bin_edges = c(0,20,40,60,80)
> data_bins = cut(heart$Age,breaks = bin_edges,include.lowest = TRUE,
+                 labels =  c("Bin1","Bin2","Bin3","Bin4"))
> bin_counts  = table(data_bins)
> print(bin_counts)
data_bins
Bin1 Bin2 Bin3 Bin4
 381 2450 2379 2331
> |
```

```r
# Feature  Scaling using r  - NORMALIZATION
heart$BMI<- (sapply(heart$BMI, function(x)
  (heart$BMI-mean(heart$BMI))/sd(heart$BMI)))
heart$Cholesterol <-(sapply(df, function(x)
  (heart$Cholesterol-mean(heart$Cholesterol))/sd(heart$Cholesterol)))
heart

```

```
 BMI.3786      BMI.3787      BMI.3788      BMI.3789      BMI.3790
 BMI.3791      BMI.3792      BMI.3793      BMI.3794      BMI.3795
 BMI.3796      BMI.3797      BMI.3798      BMI.3799      BMI.3800
 BMI.3801      BMI.3802      BMI.3803      BMI.3804      BMI.3805
 BMI.3806      BMI.3807      BMI.3808      BMI.3809      BMI.3810
 BMI.3811      BMI.3812      BMI.3813      BMI.3814      BMI.3815
 BMI.3816      BMI.3817      BMI.3818      BMI.3819      BMI.3820
 BMI.3821      BMI.3822      BMI.3823      BMI.3824      BMI.3825
```

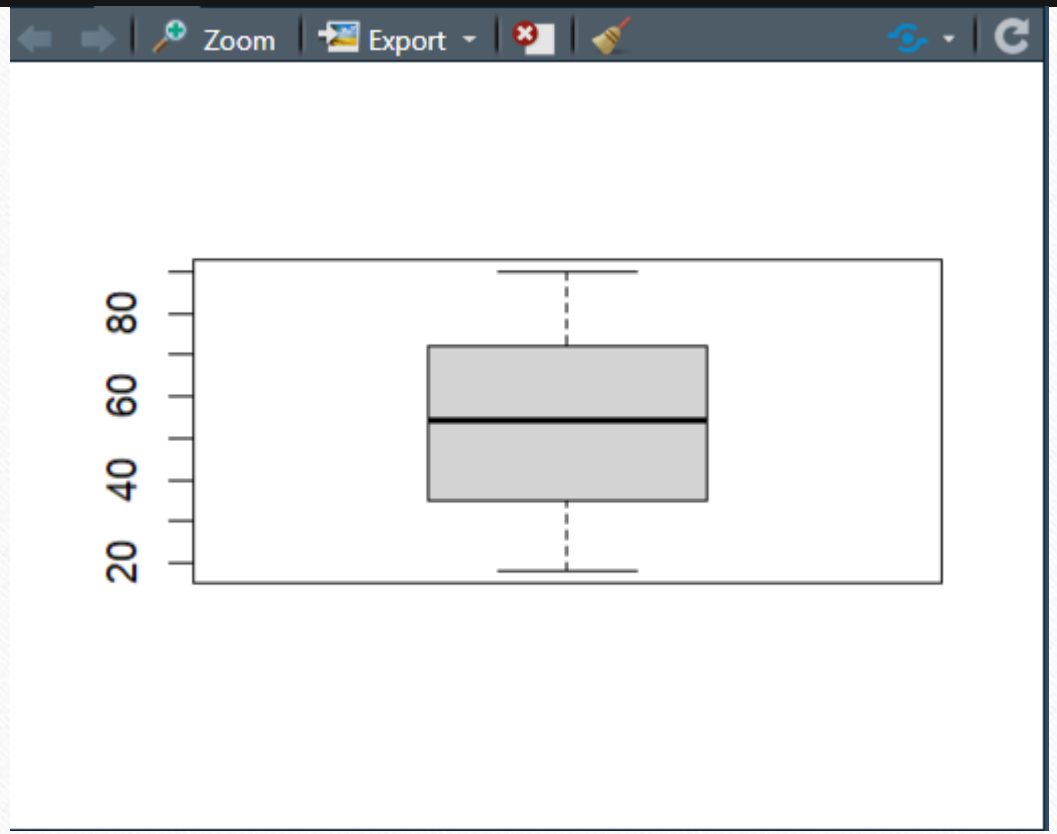
```r

# feature selection
heart_extracted = data.frame(heart$Age,heart$Sex,heart$Cholesterol,
                             heart$Blood.Pressure,heart$Diabetes,
                             heart$Obesity,heart$BMI,heart$Smoking,
                             heart$Alcohol.Consumption,
                             heart$Heart.Attack.Risk)
heart_extracted
```

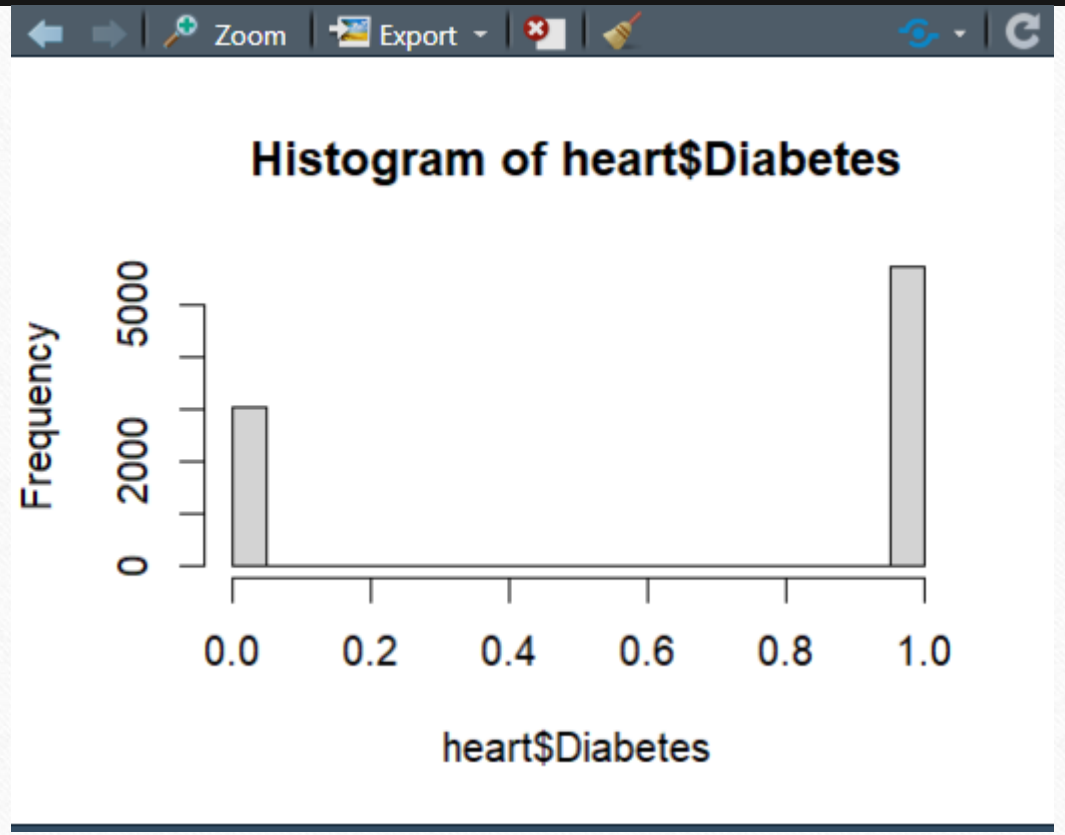
```
7          90      Male                358                 102/73
8          84      Male                220                 131/68
9          20      Male                145                144/105
10         43    Female                248                 160/70
11         73    Female                373                 107/69
12         71      Male                374                 158/71
13         45      Male                228                 101/72
14         60      Male                259                 169/72
15         88      Male                297                 112/81
16         73      Male                122                 114/88
100        28      Male                276                  92/71
    heart.Diabetes heart.Obesity heart.BMI heart.Smoking
1                0             0  31.25123             1
2                1             1  27.19497             1
3                1             0  28.17657             0
4                1             0  36.46470             1
5                1             1  21.80914             1
6                1             0  20.14684             1
7                0             0  28.88581             1
8                0             1  22.22186             1
    heart.Alcohol.Consumption heart.Heart.Attack.Risk
1                           0                       0
2                           1                       0
3                           0                       0
4                           1                       0
5                           0                       0
```

```
58  # Correlation
59  result = cor(heart$Age, heart$Cholesterol, method = "pearson")
60  print(result)
61
62  result1 = cor(heart$Cholesterol, heart$Heart.Attack.Risk, method = "pear
63  print(result1)
64
> # Correlation
> result = cor(heart$Age, heart$Cholesterol, method = "pearson")
> print(result)
[1] NA
> result1 = cor(heart$Cholesterol, heart$Heart.Attack.Risk, method = "pearso
n")
> print(result1)
[1] 0.01933968
>
```

```
65  # Outlier Detection
66  boxplot(heart$Age , ylabel = " Age")
67
```



```
67
68 hist(heart$Diabetes , ylabel ="Diabetes")
69
```

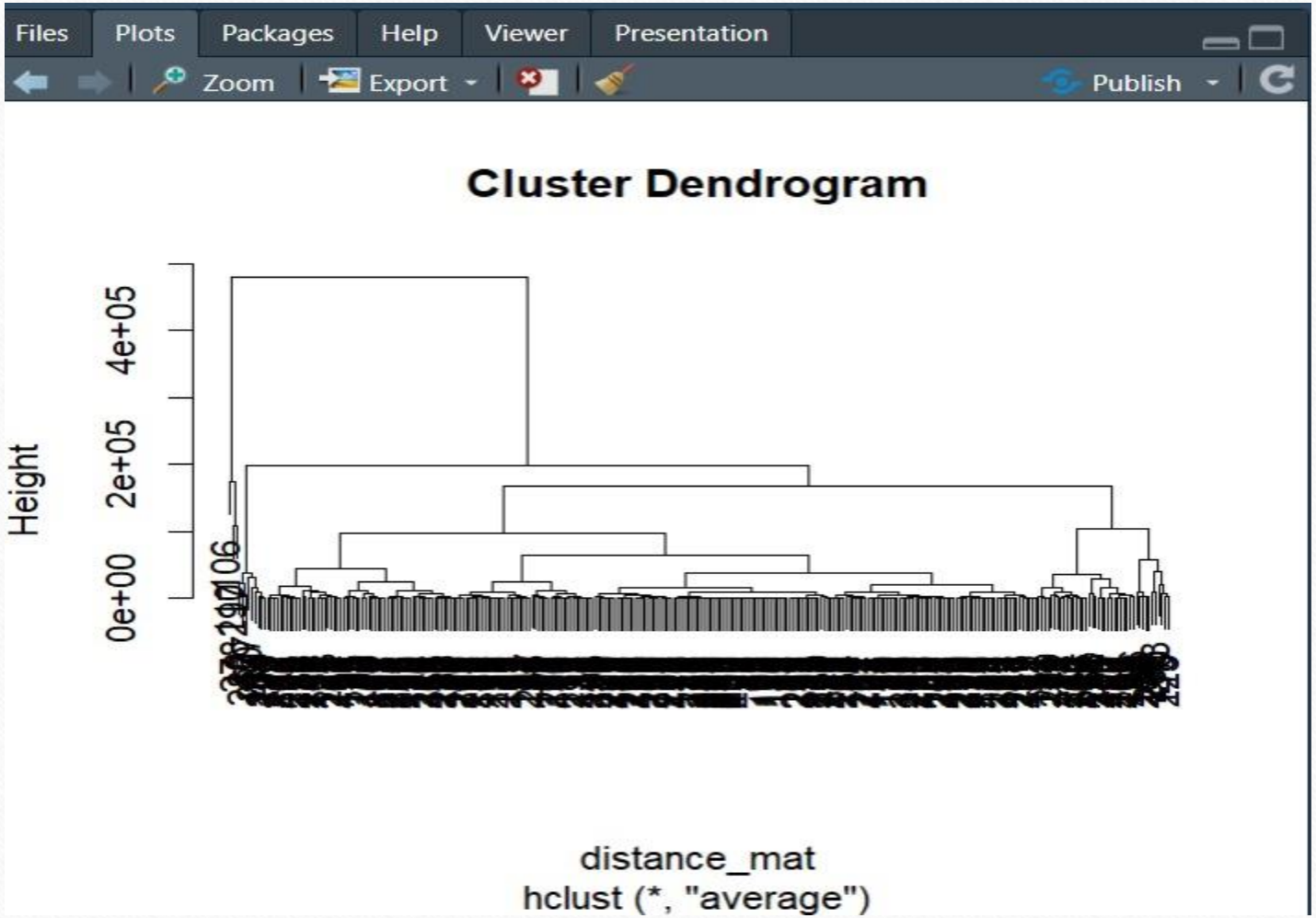


```r

install.packages("dplyr")
library(dplyr)
head(hf)
distance_mat <- dist(hf, method = 'euclidean')
distance_mat
set.seed(240)
Hierar_cl <- hclust(distance_mat, method = "average")
Hierar_cl
plot(Hierar_cl)

```

```
> distance_mat
                1             2             3             4             5             6             7
2   1.642475e+03
3   1.030000e+05 1.013580e+05
4   5.500001e+04 5.335804e+04 4.800000e+04
                8             9            10            11            12            13            14
2
3
4
               15            16            17            18            19            20            21
2
3
4
               22            23            24            25            26            27            28
2
3
4
               29            30            31            32            33            34            35
2
3
```


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Cluster Dendrogram
Height
0e+00
2e+05
4e+05
distance_mat
hclust (*, "average")

CONCLUSION

VARIOUS INSIGHTS FROM THE DATA MINING PROCESS SUCH AS FINDING THE NULL VALUES, REPLACING THE NULL VALUES, EXTRACTING THE DESCRIPTIVE STATISTICS FROM THE DATASET, BINNING,ETC. ARE PERFORMED TO KNOW ABOUT THE DATASET.