Project Title: Applying Data Pre-processing on a Dataset

Project Overview:

The given dataset contains statistics in dataset1 per 100,000 residents for assault and murder, in each of the 50 US states, in 1973. Also given is the percentage of the population living in urban areas. We have to apply the pre-processing techniques to prepare the dataset for data analysis. To prepare a cleaned dataset, we have to perform the following tasks of data pre-processing using R language-

1. Data cleaning

2. Data Integration

3. Data Transformation

4. Data Reduction

5. Data Discretization

After performing this step, we will have a process dataset ready to use.

Project Solution Design:

At first, we have to Create the dataset as CSV file. Then we have to import the dataset into RStudio so that we can perform the data processing operations using R language. We have to start with the Data cleaning process. Here we will clean the data such as we have to deal with missing values or smooth noisy data. Then comes data Integration, here we integrated a column named “Population_level” into the dataset. Then we have to do transformation. Such as converting values into numerical or other types. After that data reduction here, we reduced the data which is unnecessary. And in data discretization we have to make the data set discrete.

Data pre-processing:

1.Data cleaning:

At first, we have handled missing data in the given dataset. In the City Column we can see for the City Georgia there is no assault data. So, this is shown as NA in dataset. Here we handled the missing data by replacing NA in column “Assault”

With the mean of remaining values in the column.

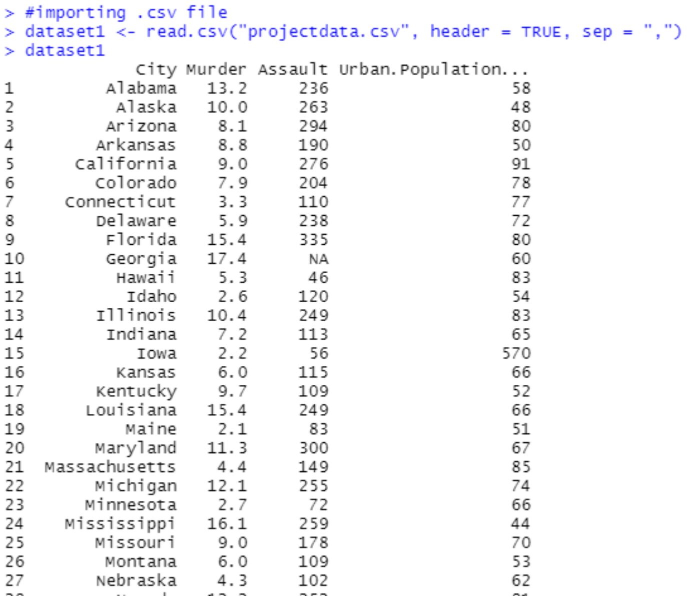
RCODE: Replacing NA in column Assault with mean of the remaining values:

#Replacing NA in column Assault with mean of the remaining values

dataset1$Assault[is.na(dataset1$Assault)]<-mean(dataset1$Assault,na.rm=TRUE)

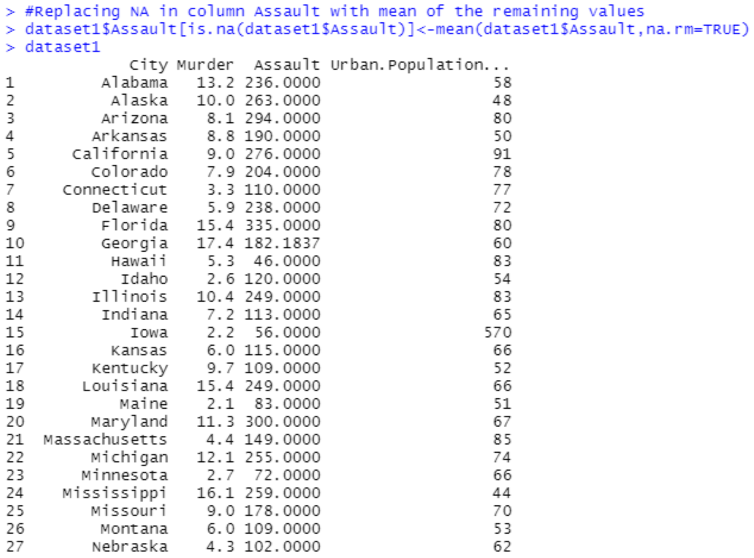
dataset1

After loading the CSV file missing values are replaced by NA:



```
> #importing .csv file
> dataset1 <- read.csv("projectdata.csv", header = TRUE, sep = ",")
> dataset1
            City Murder Assault Urban.Population...
1        Alabama   13.2     236                  58
2         Alaska   10.0     263                  48
3        Arizona    8.1     294                  80
4       Arkansas    8.8     190                  50
5     California    9.0     276                  91
6       Colorado    7.9     204                  78
7    Connecticut    3.3     110                  77
8       Delaware    5.9     238                  72
9        Florida   15.4     335                  80
10       Georgia   17.4      NA                  60
11        Hawaii    5.3      46                  83
12         Idaho    2.6     120                  54
13      Illinois   10.4     249                  83
14       Indiana    7.2     113                  65
15          Iowa    2.2      56                 570
16        Kansas    6.0     115                  66
17      Kentucky    9.7     109                  52
18     Louisiana   15.4     249                  66
19         Maine    2.1      83                  51
20      Maryland   11.3     300                  67
21 Massachusetts    4.4     149                  85
22      Michigan   12.1     255                  74
23     Minnesota    2.7      72                  66
24   Mississippi   16.1     259                  44
25      Missouri    9.0     178                  70
26       Montana    6.0     109                  53
27      Nebraska    4.3     102                  62
```

After Replacing NA in column Assault with mean of the remaining values:



```
> #Replacing NA in column Assault with mean of the remaining values
> dataset1$Assault[is.na(dataset1$Assault)]<-mean(dataset1$Assault,na.rm=TRUE)
> dataset1
            City Murder  Assault Urban.Population...
1        Alabama   13.2 236.0000                  58
2         Alaska   10.0 263.0000                  48
3        Arizona    8.1 294.0000                  80
4       Arkansas    8.8 190.0000                  50
5     California    9.0 276.0000                  91
6       Colorado    7.9 204.0000                  78
7    Connecticut    3.3 110.0000                  77
8       Delaware    5.9 238.0000                  72
9        Florida   15.4 335.0000                  80
10       Georgia   17.4 182.1837                  60
11        Hawaii    5.3  46.0000                  83
12         Idaho    2.6 120.0000                  54
13      Illinois   10.4 249.0000                  83
14       Indiana    7.2 113.0000                  65
15          Iowa    2.2  56.0000                 570
16        Kansas    6.0 115.0000                  66
17      Kentucky    9.7 109.0000                  52
18     Louisiana   15.4 249.0000                  66
19         Maine    2.1  83.0000                  51
20      Maryland   11.3 300.0000                  67
21 Massachusetts    4.4 149.0000                  85
22      Michigan   12.1 255.0000                  74
23     Minnesota    2.7  72.0000                  66
24   Mississippi   16.1 259.0000                  44
25      Missouri    9.0 178.0000                  70
26       Montana    6.0 109.0000                  53
27      Nebraska    4.3 102.0000                  62
```

We can see that data 10 was replaced by average value.

2. <u>Data Transformation:</u>

In This step we transformed the column Murder and Assault in as Numeric value.

After replacing the NA value, we can see the column has now 4-digit decimal values. So, we have to format that as numeric.

For Murder column murder cannot be of fraction so we formatted it to numeric.

RCODE: #Data Formatting... To round up the Murder and Assault variable

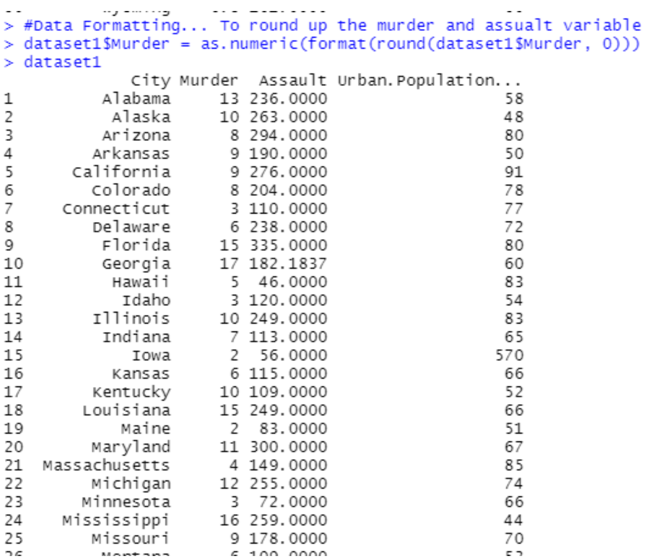
#Data Formatting... To round up the murder and assualt variable

dataset1$Murder = as.numeric(format(round(dataset1$Murder, 0)))

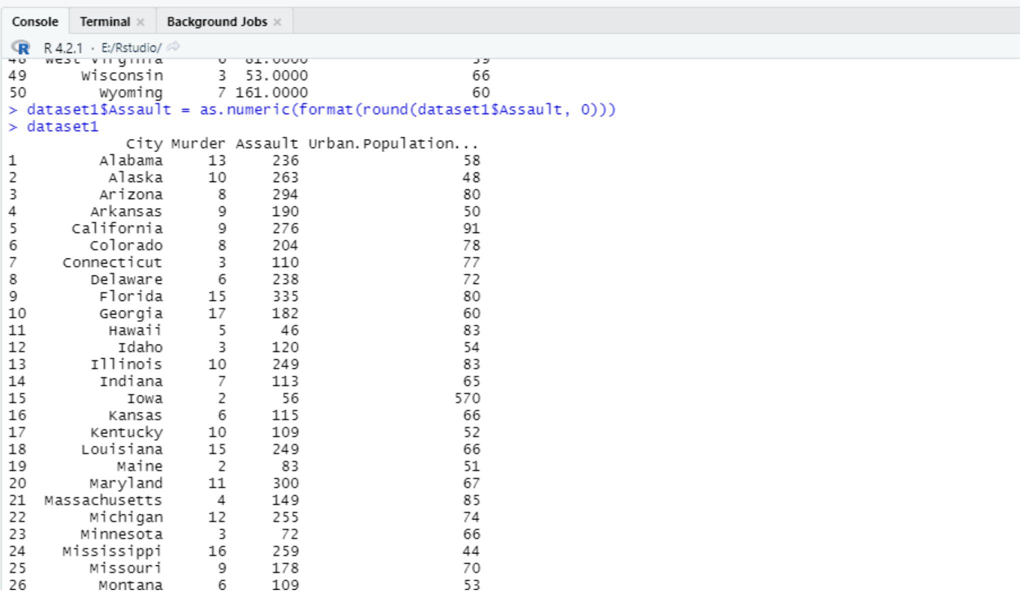
dataset1

dataset1$Assault = as.numeric(format(round(dataset1$Assault, 0)))

dataset1arrest



```
> #Data Formatting... To round up the murder and assualt variable
> dataset1$Murder = as.numeric(format(round(dataset1$Murder, 0)))
> dataset1
             City Murder  Assault Urban.Population...
1         Alabama     13 236.0000                  58
2          Alaska     10 263.0000                  48
3         Arizona      8 294.0000                  80
4        Arkansas      9 190.0000                  50
5      California      9 276.0000                  91
6        Colorado      8 204.0000                  78
7     Connecticut      3 110.0000                  77
8        Delaware      6 238.0000                  72
9         Florida     15 335.0000                  80
10        Georgia     17 182.1837                  60
11         Hawaii      5  46.0000                  83
12          Idaho      3 120.0000                  54
13       Illinois     10 249.0000                  83
14        Indiana      7 113.0000                  65
15           Iowa      2  56.0000                 570
16         Kansas      6 115.0000                  66
17       Kentucky     10 109.0000                  52
18      Louisiana     15 249.0000                  66
19          Maine      2  83.0000                  51
20       Maryland     11 300.0000                  67
21  Massachusetts      4 149.0000                  85
22       Michigan     12 255.0000                  74
23      Minnesota      3  72.0000                  66
24    Mississippi     16 259.0000                  44
25       Missouri      9 178.0000                  70
```



```
48   West Virginia      6  81.0000                  39
49       Wisconsin      3  53.0000                  66
50         Wyoming      7 161.0000                  60
> dataset1$Assault = as.numeric(format(round(dataset1$Assault, 0)))
> dataset1
             City Murder Assault Urban.Population...
1         Alabama     13     236                  58
2          Alaska     10     263                  48
3         Arizona      8     294                  80
4        Arkansas      9     190                  50
5      California      9     276                  91
6        Colorado      8     204                  78
7     Connecticut      3     110                  77
8        Delaware      6     238                  72
9         Florida     15     335                  80
10        Georgia     17     182                  60
11         Hawaii      5      46                  83
12          Idaho      3     120                  54
13       Illinois     10     249                  83
14        Indiana      7     113                  65
15           Iowa      2      56                 570
16         Kansas      6     115                  66
17       Kentucky     10     109                  52
18      Louisiana     15     249                  66
19          Maine      2      83                  51
20       Maryland     11     300                  67
21  Massachusetts      4     149                  85
22       Michigan     12     255                  74
23      Minnesota      3      72                  66
24    Mississippi     16     259                  44
25       Missouri      9     178                  70
26        Montana      6     109                  53
```

After data transformation the dataset looks like this.

3.Data Integration:

At first, we created a duplicate dataset dataset2 based on given dataset arrest because I want to keep the original dataset as it is. Then I created a new column named “Population_level” and integrated it in the data set.

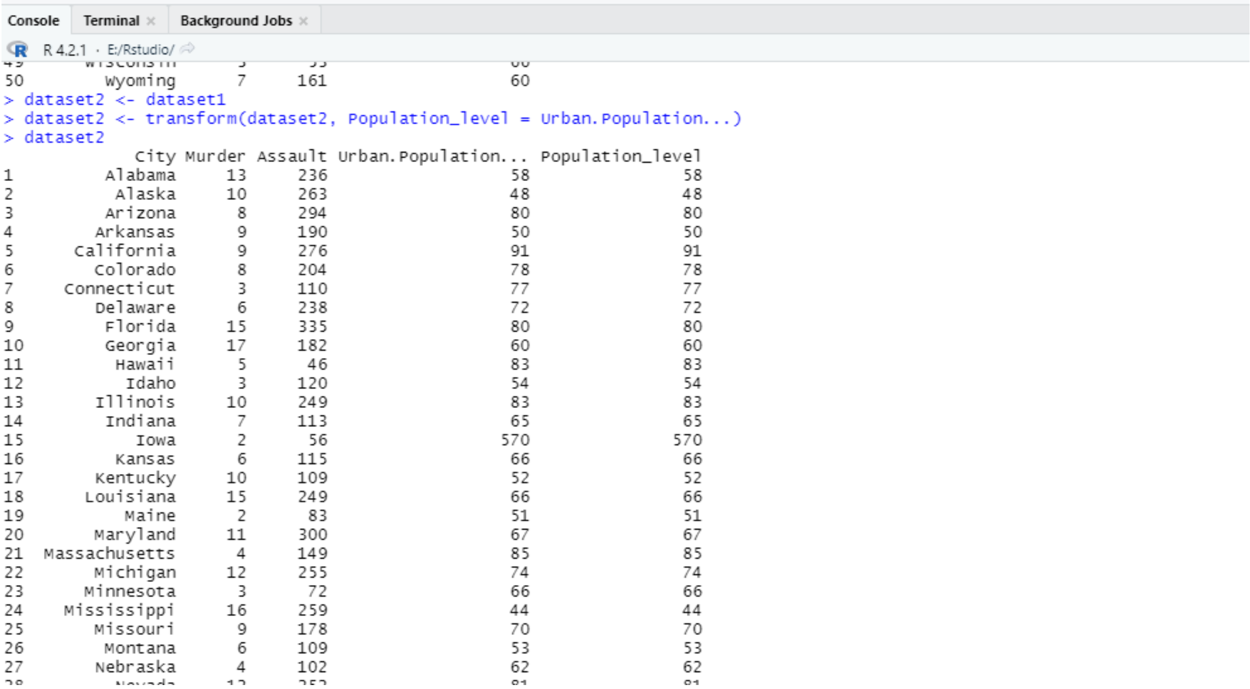
As the requirement was to Convert the urban population percentage into population_level, Such as-
small (<50%), medium (<60%), large (<70%), extra-large (<70% and above)

For that we have used conditional statement (IF-ELSE) and then used “sapply” to to store the converted data to new column “Population_level”.

RCODE: Merging Population_level variable in dataset:

```
dataset2 <- dataset1

dataset2 <- transform(dataset2, Type = Urban.Population...)

dataset2
```

Population_level column Created using Urban Population Column:



```
50        Wyoming      7     161                   60
> dataset2 <- dataset1
> dataset2 <- transform(dataset2, Population_level = Urban.Population...)
> dataset2
             City Murder Assault Urban.Population... Population_level
1         Alabama     13     236                  58               58
2          Alaska     10     263                  48               48
3         Arizona      8     294                  80               80
4        Arkansas      9     190                  50               50
5      California      9     276                  91               91
6        Colorado      8     204                  78               78
7     Connecticut      3     110                  77               77
8        Delaware      6     238                  72               72
9         Florida     15     335                  80               80
10        Georgia     17     182                  60               60
11         Hawaii      5      46                  83               83
12          Idaho      3     120                  54               54
13       Illinois     10     249                  83               83
14        Indiana      7     113                  65               65
15           Iowa      2      56                 570              570
16         Kansas      6     115                  66               66
17       Kentucky     10     109                  52               52
18      Louisiana     15     249                  66               66
19          Maine      2      83                  51               51
20       Maryland     11     300                  67               67
21  Massachusetts      4     149                  85               85
22       Michigan     12     255                  74               74
23      Minnesota      3      72                  66               66
24    Mississippi     16     259                  44               44
25       Missouri      9     178                  70               70
26        Montana      6     109                  53               53
27       Nebraska      4     102                  62               62
```

RCODE: Data integration prepare the dataset to integrate a new column (named Population_level) based on the urban population variable:

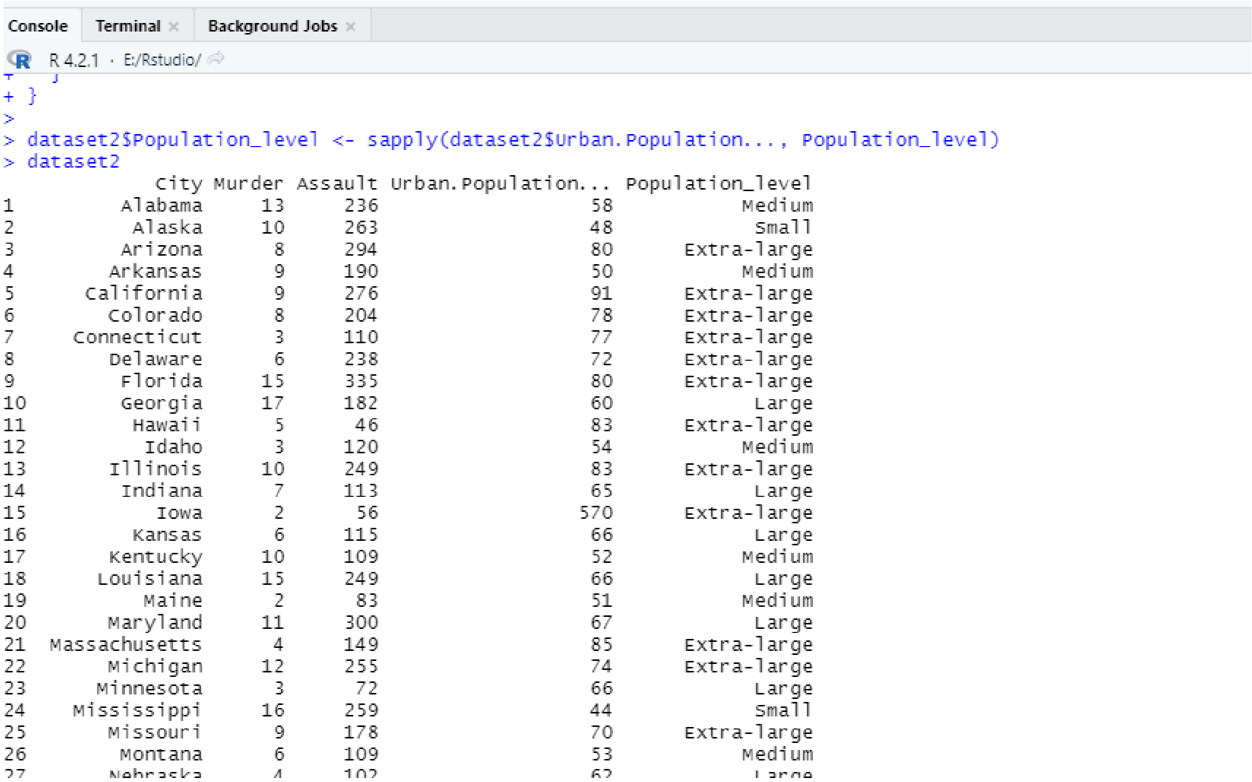
```
#Data intrigation  prepare the dataset to integrate a new column (named population_level) based on the urban population variable.

Population_level <- function(Urban.Population...){

 if (Urban.Population... < 50) {

  return("Small")

 } else if (Urban.Population... >= 50 & Urban.Population... < 60) {

  return("Medium")

 } else if (Urban.Population... >= 60 & Urban.Population... < 70) {

  return("Large")

 } else {

  return("Extra-large")

 }
```

}

dataset2$Population_level <- sapply(dataset2$Urban.Population..., Population_level)

dataset2

After Mutating the Population_level column based on the Conditions Given:



```
+   }
+ }
>
> dataset2$Population_level <- sapply(dataset2$Urban.Population..., Population_level)
> dataset2
            City Murder Assault Urban.Population... Population_level
1        Alabama     13     236                  58           Medium
2         Alaska     10     263                  48            Small
3        Arizona      8     294                  80      Extra-large
4       Arkansas      9     190                  50           Medium
5     California      9     276                  91      Extra-large
6       Colorado      8     204                  78      Extra-large
7    Connecticut      3     110                  77      Extra-large
8       Delaware      6     238                  72      Extra-large
9        Florida     15     335                  80      Extra-large
10       Georgia     17     182                  60            Large
11        Hawaii      5      46                  83      Extra-large
12         Idaho      3     120                  54           Medium
13      Illinois     10     249                  83      Extra-large
14       Indiana      7     113                  65            Large
15          Iowa      2      56                 570      Extra-large
16        Kansas      6     115                  66            Large
17      Kentucky     10     109                  52           Medium
18     Louisiana     15     249                  66            Large
19         Maine      2      83                  51           Medium
20      Maryland     11     300                  67            Large
21 Massachusetts      4     149                  85      Extra-large
22      Michigan     12     255                  74      Extra-large
23     Minnesota      3      72                  66            Large
24   Mississippi     16     259                  44            Small
25      Missouri      9     178                  70      Extra-large
26       Montana      6     109                  53           Medium
27      Nebraska      4     102                  62            Large
```

4.Data Reduction:

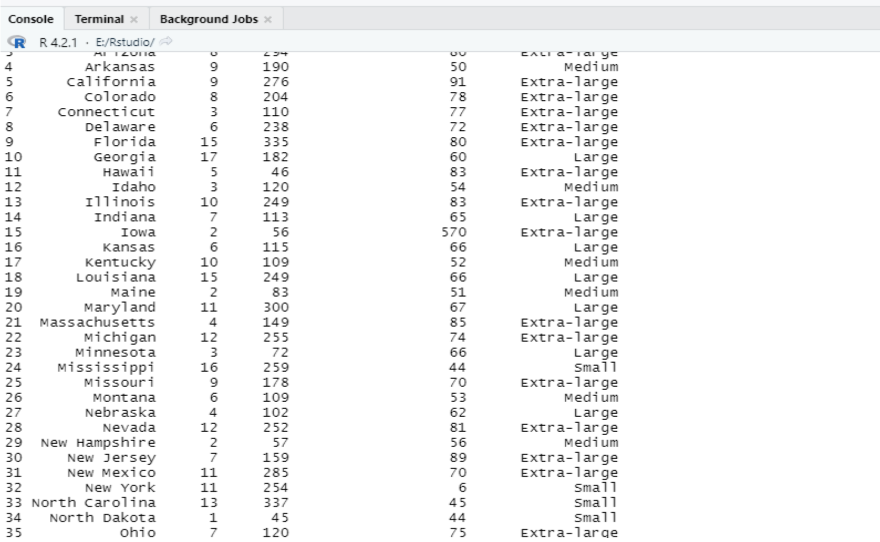
In this step we have removed 2 rows from the dataset.
As nation 15 is too high to be a percentage value and nation 32 is too low to be a percentage, we decided to remove the value for better data processing.

RCODE: Data reduction (as nation 15 is too high and nation 32 is too low to be a percentage):

dataset2 <- dataset2[-c(15, 32),]

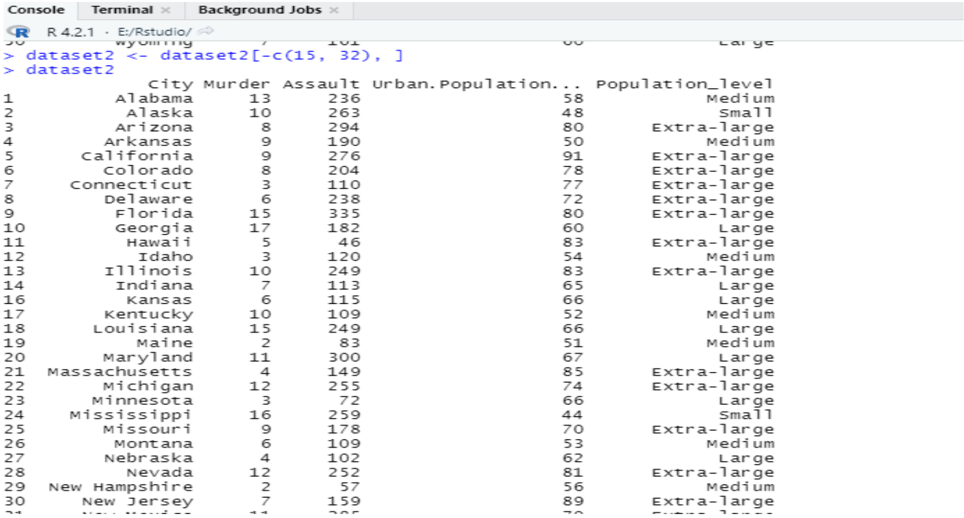
dataset2

Before Reduction:



```
4         Arkansas   9     190                    50      Medium
5       California   9     276                    91 Extra-large
6         Colorado   8     204                    78 Extra-large
7      Connecticut   3     110                    77 Extra-large
8         Delaware   6     238                    72 Extra-large
9          Florida  15     335                    80 Extra-large
10         Georgia  17     182                    60       Large
11          Hawaii   5      46                    83 Extra-large
12           Idaho   3     120                    54      Medium
13        Illinois  10     249                    83 Extra-large
14         Indiana   7     113                    65       Large
15            Iowa   2      56                   570 Extra-large
16          Kansas   6     115                    66       Large
17        Kentucky  10     109                    52      Medium
18       Louisiana  15     249                    66       Large
19           Maine   2      83                    51      Medium
20        Maryland  11     300                    67       Large
21   Massachusetts   4     149                    85 Extra-large
22        Michigan  12     255                    74 Extra-large
23       Minnesota   3      72                    66       Large
24     Mississippi  16     259                    44       Small
25        Missouri   9     178                    70 Extra-large
26         Montana   6     109                    53      Medium
27        Nebraska   4     102                    62       Large
28          Nevada  12     252                    81 Extra-large
29   New Hampshire   2      57                    56      Medium
30      New Jersey   7     159                    89 Extra-large
31      New Mexico  11     285                    70 Extra-large
32        New York  11     254                     6       Small
33  North Carolina  13     337                    45       Small
34    North Dakota   1      45                    44       Small
35            Ohio   7     120                    75 Extra-large
```

After Reduction:



```
> dataset2 <- dataset2[-c(15, 32), ]
> dataset2
             City Murder Assault Urban.Population... Population_level
1         Alabama     13     236                  58           Medium
2          Alaska     10     263                  48            Small
3         Arizona      8     294                  80      Extra-large
4        Arkansas      9     190                  50           Medium
5      California      9     276                  91      Extra-large
6        Colorado      8     204                  78      Extra-large
7     Connecticut      3     110                  77      Extra-large
8        Delaware      6     238                  72      Extra-large
9         Florida     15     335                  80      Extra-large
10        Georgia     17     182                  60            Large
11         Hawaii      5      46                  83      Extra-large
12          Idaho      3     120                  54           Medium
13       Illinois     10     249                  83      Extra-large
14        Indiana      7     113                  65            Large
16         Kansas      6     115                  66            Large
17       Kentucky     10     109                  52           Medium
18      Louisiana     15     249                  66            Large
19          Maine      2      83                  51           Medium
20       Maryland     11     300                  67            Large
21  Massachusetts      4     149                  85      Extra-large
22       Michigan     12     255                  74      Extra-large
23      Minnesota      3      72                  66            Large
24    Mississippi     16     259                  44            Small
25       Missouri      9     178                  70      Extra-large
26        Montana      6     109                  53           Medium
27       Nebraska      4     102                  62            Large
28         Nevada     12     252                  81      Extra-large
29  New Hampshire      2      57                  56           Medium
30     New Jersey      7     159                  89      Extra-large
```

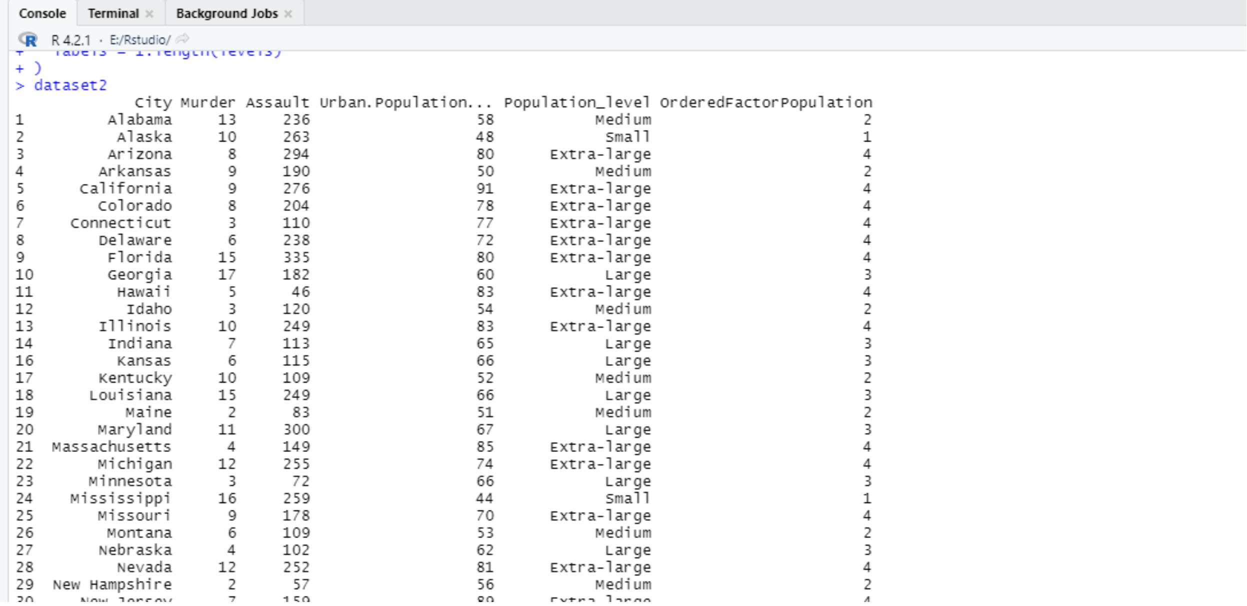
As we can see number 15 and 32 is removed.

Integrate new column “OrderedFactorPopulation” like (Small=1,Medium=2,Large=3,Extra-large=4):

levels <- c("Small", "Medium", "Large", "Extra-large")

```
dataset2$OrderedFactorPopulation <- factor(
  dataset2$Population_level,
  levels = levels,
  ordered = TRUE,
  labels = 1:length(levels)
)
dataset2
```

After Integrating new column “OrderedFactorPopulation”:



```
+   labels = 1:length(levels)
+ )
> dataset2
            City Murder Assault Urban.Population... Population_level OrderedFactorPopulation
1        Alabama     13     236                  58           Medium                       2
2         Alaska     10     263                  48            Small                       1
3        Arizona      8     294                  80      Extra-large                       4
4       Arkansas      9     190                  50           Medium                       2
5     California      9     276                  91      Extra-large                       4
6       Colorado      8     204                  78      Extra-large                       4
7    Connecticut      3     110                  77      Extra-large                       4
8       Delaware      6     238                  72      Extra-large                       4
9        Florida     15     335                  80      Extra-large                       4
10       Georgia     17     182                  60            Large                       3
11        Hawaii      5      46                  83      Extra-large                       4
12         Idaho      3     120                  54           Medium                       2
13      Illinois     10     249                  83      Extra-large                       4
14       Indiana      7     113                  65            Large                       3
16        Kansas      6     115                  66            Large                       3
17      Kentucky     10     109                  52           Medium                       2
18     Louisiana     15     249                  66            Large                       3
19         Maine      2      83                  51           Medium                       2
20      Maryland     11     300                  67            Large                       3
21 Massachusetts      4     149                  85      Extra-large                       4
22      Michigan     12     255                  74      Extra-large                       4
23     Minnesota      3      72                  66            Large                       3
24   Mississippi     16     259                  44            Small                       1
25      Missouri      9     178                  70      Extra-large                       4
26       Montana      6     109                  53           Medium                       2
27      Nebraska      4     102                  62            Large                       3
28        Nevada     12     252                  81      Extra-large                       4
29 New Hampshire      2      57                  56           Medium                       2
30    New Jersey      7     159                  89      Extra-large                       4
```

5.Data Discretization:

Data discretization refers to a method of converting a huge number of data values into smaller ones so that the evaluation and management of data become easy.

We did that in Step 3. Data Integration part.

Before discretization there was no type or limit to determine the urban population what portion do they belong from.

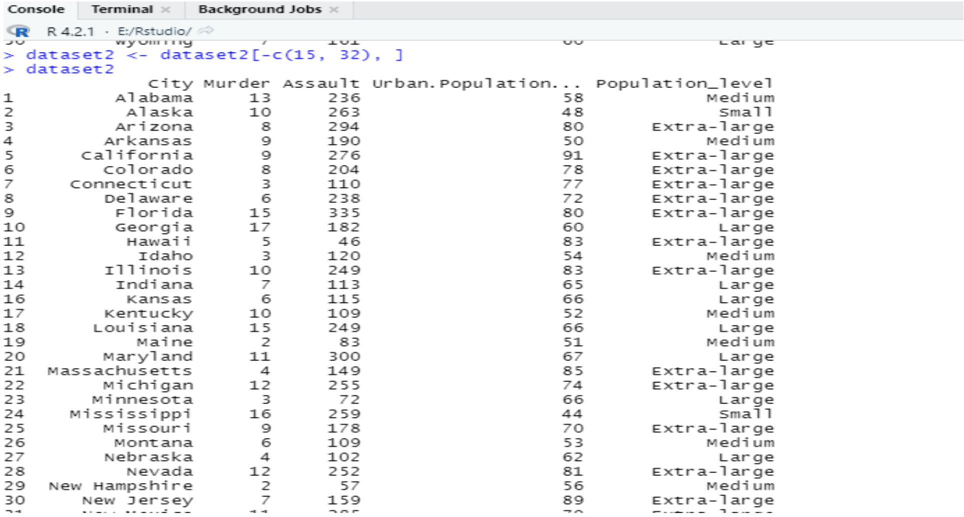
After discretization we can easily say that below 50% population belongs to small portion.

Similarly, below 60% population belongs to medium portion.

below 70% population belongs to large portion.

And from 70% and above population belongs to Extra Large portion.

After Discretization the dataset:



```
> dataset2 <- dataset2[-c(15, 32), ]
> dataset2
            City Murder Assault Urban.Population... Population_level
1        Alabama     13     236                  58           Medium
2         Alaska     10     263                  48            Small
3        Arizona      8     294                  80      Extra-large
4       Arkansas      9     190                  50           Medium
5     California      9     276                  91      Extra-large
6       Colorado      8     204                  78      Extra-large
7    Connecticut      3     110                  77      Extra-large
8       Delaware      6     238                  72      Extra-large
9        Florida     15     335                  80      Extra-large
10       Georgia     17     182                  60            Large
11        Hawaii      5      46                  83      Extra-large
12         Idaho      3     120                  54           Medium
13      Illinois     10     249                  83      Extra-large
14       Indiana      7     113                  65            Large
16        Kansas      6     115                  66            Large
17      Kentucky     10     109                  52           Medium
18     Louisiana     15     249                  66            Large
19         Maine      2      83                  51           Medium
20      Maryland     11     300                  67            Large
21 Massachusetts      4     149                  85      Extra-large
22      Michigan     12     255                  74      Extra-large
23     Minnesota      3      72                  66            Large
24   Mississippi     16     259                  44            Small
25      Missouri      9     178                  70      Extra-large
26       Montana      6     109                  53           Medium
27      Nebraska      4     102                  62            Large
28        Nevada     12     252                  81      Extra-large
29 New Hampshire      2      57                  56           Medium
30    New Jersey      7     159                  89      Extra-large
```

Discussion and Conclusion:

After doing date pre-processing operations in a given dataset, we can perform these steps in any datasets when we need. And data pre-processing helps AI or machine learning to easily analyze the data. We can also get an easy-to-understand dataset after doing these operations. It makes machines to understand a huge data easily and properly without facing any problem and errors.