Department of Computer Science
American International University-Bangladesh
Mid Term Report

Course Name: INTRODUCTION TO DATA SCIENCE

"Project on Data Pre-Processing"

Supervised By:

Dr. Akinul Islam Jony

Associate Professor, Computer Science -AIUB

Submitted By:

Sumaiya Malik

ID: 20-43688-2

Section: B

Submission Date: March 1,2023.

Project Title: Applying Data Pre-processing on a Dataset.

Project Overview:

Text, photos, video, and other types of unprocessed, real-world data are disorganized. In addition to the possibility of faults and inconsistencies, it is frequently lacking and lacks a regular, consistent design. Machines prefer to process information that is neat and orderly; they can only read data as 1s and 0s. Hence, it is simple to calculate structured data like whole numbers and percentages. But unstructured data must first be cleaned and prepared in the form of text and graphics before analysis. Data preparation is the process of altering or encoding data so that a computer can quickly and easily understand it. For a model to provide accurate and precise predictions, the algorithm must be quick to decipher the characteristics of the data. Data preparation is therefore essential for raising the general level of data quality. Data cleaning, data integration, data transformation, data reduction, and data discretization are the four main stages of data preprocessing. Filling in missing values, reducing noise in the data, resolving discrepancies, and eliminating outliers are all steps in the data pretreatment process known as "data cleaning." A data preparation phase called data integration brings together data from several sources into one, more substantial data storage, like a data warehouse. By changing the value, structure, or format of data using methods like scaling, normalization, and others, data transformation is a methodology for turning high-quality data into various formats. In order to change the data into the right form, data transformation techniques also include data cleaning and data reduction. To produce patterns that are simpler to grasp, data transformation is a crucial data preprocessing technique that must be applied to the data before data mining. This method of pre-processing data is systematic.

The following dataset of the reports contains statistics in arrests 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. In our given dataset initially, I have seen some missing value. After handle missing value it may occurs some noise like format problem. Then I use a defined format with my preferred methodology. I inserted a new column using a different column value from the one that was provided in our inquiry condition after adjusting the format. I tried to handle converting categorical values to numerical values, which is our discretization part, after adding the column. Finally, I attempted to compare the values of three columns using the normalizing technique.

Project Solution Design:

In this project, we are required to perform the techniques of Data pre-processing to obtain a clean dataset ready for Data Analysis. Here's a sequence of the project.

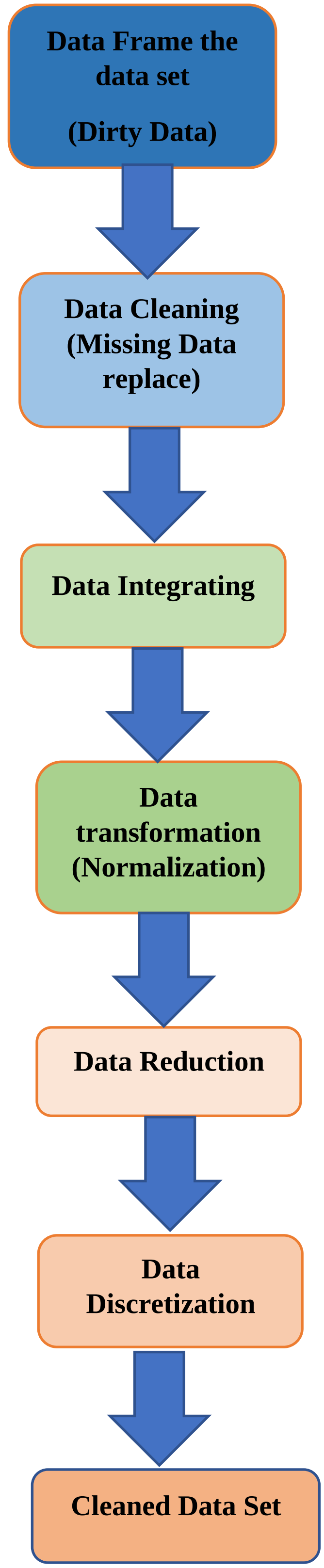


Figure 1: Block Diagram of the project Solution.

Software Used for This Project:

RStudio will be the program we employ to shape the data. A sophisticated and user-friendly Integrated Development Environment (IDE) that offers a one-stop shop for all statistical computing and graphics, RStudio is a powerful and simple method to engage with R programming. A more sophisticated version of R called The RStudio features a multi-pane window layout that gives users access to all essential features on a single screen (such as source, console, environment & history, files, photos, graphs, etc.). Then, we used a pdf to excel converter application to convert the data set to an excel file.

Data pre-processing:

1. **Data Frame dataset:** Before we start preprocessing our data, first we need to do data frame using the following code lines.

   Code:

```
> X<-c("Alabama", "Alaska","Arizona","Arkansas","California","Colorado","Connecticut"," Delaware","Florida","Georgia","Hawaii","Idaho","Illinois","Indiana","lowa","kansas","Kentucky","Louisiana","Maine","Maryland","Massachusetts","Michigan","Minnesota","Mississippi","Missouri","Montana","Nebraska","Nevada","New_Hampshire","New Jersey","New Mexico","New York","North Carolina","North Dakota","Ohio","Oklahoma","Oregon","Pennsylvania","Rhode Island","South Carolina","South Dakota","Tennessee","Texas","Utah","Vermont"," Virginia","Washington","West Virginia","Wisconsim","Wyoming")

> Murder<-c(13.2,10,8.1,8.8,9,7.9,3.3,5.9,15.4,17.4,5.3,2.6,10.4,7.2,2.2,6,9.7,15.4,2.1,11.3,4.4,12.1,2.7,16.1,9,6,4.3,12.2,2.1,7.4,11.4,11.1,13,0.8,7.3,6.6,4.9,6.3,3.4,14.4,3.8,13.2,12.7,3.2,2.2,8.5,4,5.7,2.6,6.8)

> Assault<-c(236,263,294,190,276,204,110,238,335,0,46,120,249,11,56,115,109,249,83,300,149,255,72,259,178,109,102,252,57,159,285,254,337,45,120,151,159,106,174,879,86,188,201,120,48,156,145,81,53,161)

> Urban.Population<-c(58,48,80,50,91,78,77,72,80,60,83,54,83,65,570,66,52,66,51,67,85,74,66,44,70,53,62,81,56,89,70,6,45,44,75,68,67,72,87,48,45,59,80,80,32,63,73,39,66,60)

> dataset<-data.frame(X,Murder,Assault,Urban.Population)

> dataset
```

```
R 4.2.2 · ~/
4,11.4,11.1,13,0.8,7.3,6.6,4.9,6.3,3.4,14.4,3.8,13.2,12.7,3.2,2.
> Assault<-c(236,263,294,190,276,204,110,238,335,0,46,120,249,11
54,337,45,120,151,159,106,174,879,86,188,201,120,48,156,145,81,5
> UrbanPopulation<-c(58,48,80,50,91,78,77,72,80,60,83,54,83,65,5
7,72,87,48,45,59,80,80,32,63,73,39,66,60)
> dataset<-data.frame(X,Murder,Assault,UrbanPopulation)
> dataset
                X Murder Assault UrbanPopulation
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       0              60
11         Hawaii    5.3      46              83
12          Idaho    2.6     120              54
13       Illinois   10.4     249              83
14        Indiana    7.2      11              65
15           lowa    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
28         Nevada   12.2     252              81
29  New_Hampshire    2.1      57              56
30     New Jersey    7.4     159              89
31     New Mexico   11.4     285              70
32       New York   11.1     254               6
33 North Carolina   13.0     337              45
34   North Dakota    0.8      45              44
35           Ohio    7.3     120              75
```

Figure 2: After performing the Data Frame.

2.Data Cleaning:

2.1 Data Munging: Since in this dataset all the data are per 100,000 residents, there are no data munging steps in the data set.

2.2Handling the missing data: The Assault column value in the dataset has missing data. This issue must be resolved before to incorporating a data set into a model; otherwise, it will seriously impact that model. So, we should handle this data set. we can handle 2 ways either replace the data or Discard. As it has only 50 data so we couldn't remove any data. so, we can replace using average function as a numerical data.

Code:

> dataset$ Assault = ifelse(is.na(dataset$ Assault), ave(dataset$ Assault, FUN= function(x)mean(x, na.rm = TRUE)), dataset$Assault)

> dataset

```
Console  Terminal  Background Jobs
R 4.2.2 · ~/
> dataset$ Assault = ifelse(is.na(dataset$ Assault), ave(dataset$ Assault, FUN = function(x)mean(x, na.rm = TRUE)), dataset$Assault)
> dataset
                X 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   17.4 169.9388                  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                  57
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
28         Nevada   12.2 252.0000                  81
29  New Hampshire    2.1  57.0000                  56
30     New Jersey    7.4 159.0000                  89
31     New Mexico   11.4 285.0000                  70
32       New York   11.1 254.0000                  86
33 North Carolina   13.0 337.0000                  45
34   North Dakota    0.8  45.0000                  44
35           Ohio    7.3 120.0000                  75
36       Oklahoma    6.6 151.0000                  68
37         Oregon    4.9 159.0000                  67
38   Pennsylvania    6.3 106.0000                  72
39   Rhode Island    3.4 174.0000                  87
40 South Carolina   14.4 279.0000                  48
41   South Dakota    3.8  86.0000                  45
42      Tennessee   13.2 188.0000                  59
43          Texas   12.7 201.0000                  80
44           Utah    3.2 120.0000                  80
45        Vermont    2.2  48.0000                  32
46       Virginia    8.5 156.0000                  63
47     Washington    4.0 145.0000                  73
48  West Virginia    5.7  81.0000                  39
```

Figure 3: After replace missing value.

2.3 <u>Smooth Noisy Data:</u> Here we first search for any outliers in the data set. Now we run the codes.

<u>Code:</u>

> outMurder <- dataset[(dataset$Murder > 20 | dataset$Murder < 1),]

> outMurder

```
10 Georgia      17.4    182.                     60
# … with 40 more rows
# ℹ Use `print(n = ...)` to see more rows
> outMurder <-  dataset[(dataset$Murder > 20 |  dataset$Murder < 1),]
> outMurder
# A tibble: 1 x 4
  States       Murder Assault `Urban population (%)`
  <chr>         <dbl>   <dbl>                  <dbl>
1 North Dakota    0.8      45                     44
> 
```

<u>Code:</u>

> outAssault <- dataset[(data$Assault >400 | dataset$Assault < 45),]

> outAssault

```
# A tibble: 1 x 4
  States         Murder Assault `Urban population (%)`
  <chr>           <dbl>   <dbl>                  <dbl>
1 South Carolina   14.4     879                     48
> 
```

<u>Code:</u>

> outUP <- dataset[(dataset$`Urban.Population (%)` < 32 | dataset$`Urban.population (%)` > 91),]

> outUP

```
> outUP <- dataset[(dataset$`Urban population (%)` < 32 |  dataset$`Urban population (%)` > 91),]
> outUP
# A tibble: 2 x 4
  States   Murder Assault `Urban population (%)`
  <chr>     <dbl>   <dbl>                  <dbl>
1 Iowa        2.2      56                    570
2 New York   11.1     254                      6
> 
```

Figure 4: After smoothing noisy data values.

3.Data Integration: Data integration is a process where we need to integrate new data from different source or table Let's now imagine that we must add a new column of data to our data table based on estimates of the urban population.

Code:

\> dataset$PopulationLevel <- rep(NA, nrow(dataset))

\> dataset

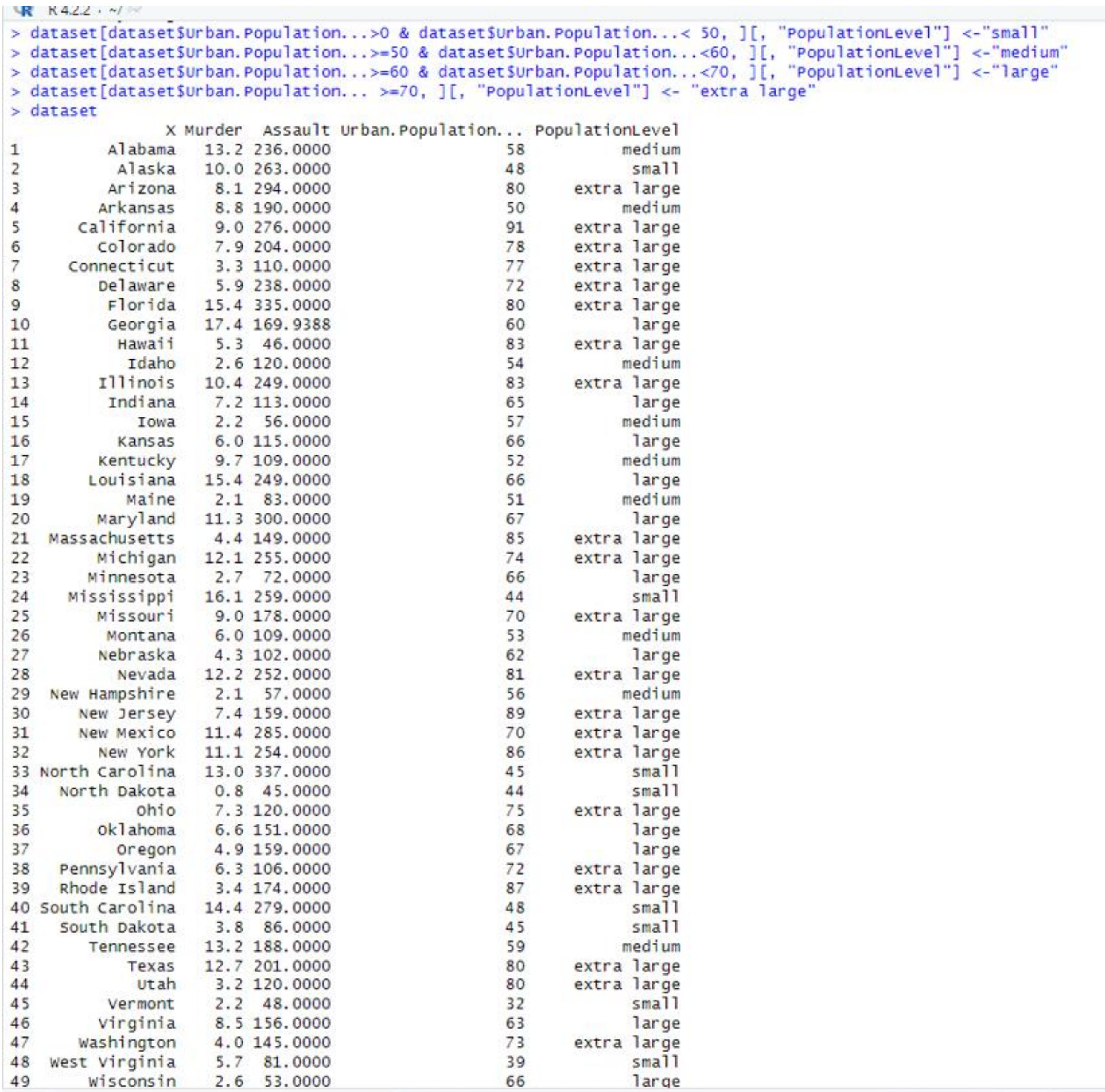
```
R 4.2.2 · ~/
> dataset$PopulationLevel <- rep(NA, nrow(dataset))
> dataset
                X Murder  Assault Urban.Population... PopulationLevel
1         Alabama   13.2 236.0000                  58              NA
2          Alaska   10.0 263.0000                  48              NA
3         Arizona    8.1 294.0000                  80              NA
4        Arkansas    8.8 190.0000                  50              NA
5      California    9.0 276.0000                  91              NA
6        Colorado    7.9 204.0000                  78              NA
7     Connecticut    3.3 110.0000                  77              NA
8        Delaware    5.9 238.0000                  72              NA
9         Florida   15.4 335.0000                  80              NA
10        Georgia   17.4 169.9388                  60              NA
11         Hawaii    5.3  46.0000                  83              NA
12          Idaho    2.6 120.0000                  54              NA
13       Illinois   10.4 249.0000                  83              NA
14        Indiana    7.2 113.0000                  65              NA
15           Iowa    2.2  56.0000                  57              NA
16         Kansas    6.0 115.0000                  66              NA
17       Kentucky    9.7 109.0000                  52              NA
18      Louisiana   15.4 249.0000                  66              NA
19          Maine    2.1  83.0000                  51              NA
20       Maryland   11.3 300.0000                  67              NA
21  Massachusetts    4.4 149.0000                  85              NA
22       Michigan   12.1 255.0000                  74              NA
23      Minnesota    2.7  72.0000                  66              NA
24    Mississippi   16.1 259.0000                  44              NA
25       Missouri    9.0 178.0000                  70              NA
26        Montana    6.0 109.0000                  53              NA
27       Nebraska    4.3 102.0000                  62              NA
28         Nevada   12.2 252.0000                  81              NA
29  New Hampshire    2.1  57.0000                  56              NA
30     New Jersey    7.4 159.0000                  89              NA
31     New Mexico   11.4 285.0000                  70              NA
32       New York   11.1 254.0000                  86              NA
33 North Carolina   13.0 337.0000                  45              NA
34   North Dakota    0.8  45.0000                  44              NA
35           Ohio    7.3 120.0000                  75              NA
36       Oklahoma    6.6 151.0000                  68              NA
37         Oregon    4.9 159.0000                  67              NA
38   Pennsylvania    6.3 106.0000                  72              NA
39   Rhode Island    3.4 174.0000                  87              NA
40 South Carolina   14.4 279.0000                  48              NA
41   South Dakota    3.8  86.0000                  45              NA
42      Tennessee   13.2 188.0000                  59              NA
43          Texas   12.7 201.0000                  80              NA
44           Utah    3.2 120.0000                  80              NA
45        Vermont    2.2  48.0000                  32              NA
46       Virginia    8.5 156.0000                  63              NA
47     Washington    4.0 145.0000                  73              NA
48  West Virginia    5.7  81.0000                  39              NA
49      Wisconsin    2.6  53.0000                  66              NA
50        Wyoming    6.8 161.0000                  60              NA
>
```

Figure 5: Adding a new column for Population Level.

Code:

> dataset[dataset$Urban.Population...>0 & dataset$Urban.Population...< 50,][, "PopulationLevel"] <-"small"
> dataset[dataset$Urban.Population...>=50 & dataset$Urban.Population...<60,][, "PopulationLevel"] <- "medium"
> dataset[dataset$Urban.Population...>=60 & dataset$Urban.Population...<70,][, "PopulationLevel"] <- "large"
> dataset[dataset$Urban.Population... >=70,][, "PopulationLevel"] <- "extra large"

>dataset



```
R 4.2.2 · ~/
> dataset[dataset$Urban.Population...>0 & dataset$Urban.Population...< 50, ][, "PopulationLevel"] <-"small"
> dataset[dataset$Urban.Population...>=50 & dataset$Urban.Population...<60, ][, "PopulationLevel"] <-"medium"
> dataset[dataset$Urban.Population...>=60 & dataset$Urban.Population...<70, ][, "PopulationLevel"] <-"large"
> dataset[dataset$Urban.Population... >=70, ][, "PopulationLevel"] <- "extra large"
> dataset
                X Murder  Assault Urban.Population... PopulationLevel
1         Alabama   13.2 236.0000                  58          medium
2          Alaska   10.0 263.0000                  48           small
3         Arizona    8.1 294.0000                  80     extra large
4        Arkansas    8.8 190.0000                  50          medium
5      California    9.0 276.0000                  91     extra large
6        Colorado    7.9 204.0000                  78     extra large
7     Connecticut    3.3 110.0000                  77     extra large
8        Delaware    5.9 238.0000                  72     extra large
9         Florida   15.4 335.0000                  80     extra large
10        Georgia   17.4 169.9388                  60           large
11         Hawaii    5.3  46.0000                  83     extra large
12          Idaho    2.6 120.0000                  54          medium
13       Illinois   10.4 249.0000                  83     extra large
14        Indiana    7.2 113.0000                  65           large
15           Iowa    2.2  56.0000                  57          medium
16         Kansas    6.0 115.0000                  66           large
17       Kentucky    9.7 109.0000                  52          medium
18      Louisiana   15.4 249.0000                  66           large
19          Maine    2.1  83.0000                  51          medium
20       Maryland   11.3 300.0000                  67           large
21  Massachusetts    4.4 149.0000                  85     extra large
22       Michigan   12.1 255.0000                  74     extra large
23      Minnesota    2.7  72.0000                  66           large
24    Mississippi   16.1 259.0000                  44           small
25       Missouri    9.0 178.0000                  70     extra large
26        Montana    6.0 109.0000                  53          medium
27       Nebraska    4.3 102.0000                  62           large
28         Nevada   12.2 252.0000                  81     extra large
29  New Hampshire    2.1  57.0000                  56          medium
30     New Jersey    7.4 159.0000                  89     extra large
31     New Mexico   11.4 285.0000                  70     extra large
32       New York   11.1 254.0000                  86     extra large
33 North Carolina   13.0 337.0000                  45           small
34   North Dakota    0.8  45.0000                  44           small
35           Ohio    7.3 120.0000                  75     extra large
36       Oklahoma    6.6 151.0000                  68           large
37         Oregon    4.9 159.0000                  67           large
38   Pennsylvania    6.3 106.0000                  72     extra large
39   Rhode Island    3.4 174.0000                  87     extra large
40 South Carolina   14.4 279.0000                  48           small
41   South Dakota    3.8  86.0000                  45           small
42      Tennessee   13.2 188.0000                  59          medium
43          Texas   12.7 201.0000                  80     extra large
44           Utah    3.2 120.0000                  80     extra large
45        Vermont    2.2  48.0000                  32           small
46       Virginia    8.5 156.0000                  63           large
47     Washington    4.0 145.0000                  73     extra large
48  West Virginia    5.7  81.0000                  39           small
49      Wisconsin    2.6  53.0000                  66           large
```

Figure 6: After Integration of a new column (Population Level).

4. Data Transformation: As has already been known, the data transformation process includes one or more of the following steps: normalization, summarization, noise removal, smoothing, and summarizing of the data. for our data set I used normalization.

4.1 Normalization:

The statistical distribution of the data is positively impacted by normalization procedures since they allow us to minimize the magnitude of the variables.in this data set I have normalized column between 2 to 4.

Code:

```
> min_max_norm <- function(x) { (x - min(x)) / (max(x) - min(x)) }
> dataset <- as.data.frame(lapply(dataset[2:4], min_max_norm))
> dataset
```

```
R 4.2.2 · ~/
> min_max_norm <- function(x) { (x - min(x)) / (max(x) - min(x)) }
> dataset <- as.data.frame(lapply(dataset[2:4], min_max_norm))
> dataset
       Murder     Assault Urban.Population...
1  0.74698795 0.654109589           0.4406780
2  0.55421687 0.746575342           0.2711864
3  0.43975904 0.852739726           0.8135593
4  0.48192771 0.496575342           0.3050847
5  0.49397590 0.791095890           1.0000000
6  0.42771084 0.544520548           0.7796610
7  0.15060241 0.222602740           0.7627119
8  0.30722892 0.660958904           0.6779661
9  0.87951807 0.993150685           0.8135593
10 1.00000000 0.427872519           0.4745763
11 0.27108434 0.003424658           0.8644068
12 0.10843373 0.256849315           0.3728814
13 0.57831325 0.698630137           0.8644068
14 0.38554217 0.232876712           0.5593220
15 0.08433735 0.037671233           0.4237288
16 0.31325301 0.239726027           0.5762712
17 0.53614458 0.219178082           0.3389831
18 0.87951807 0.698630137           0.5762712
19 0.07831325 0.130136986           0.3220339
20 0.63253012 0.873287671           0.5932203
21 0.21686747 0.356164384           0.8983051
22 0.68072289 0.719178082           0.7118644
23 0.11445783 0.092465753           0.5762712
24 0.92168675 0.732876712           0.2033898
25 0.49397590 0.455479452           0.6440678
26 0.31325301 0.219178082           0.3559322
27 0.21084337 0.195205479           0.5084746
28 0.68674699 0.708904110           0.8305085
29 0.07831325 0.041095890           0.4067797
30 0.39759036 0.390410959           0.9661017
31 0.63855422 0.821917808           0.6440678
32 0.62048193 0.715753425           0.9152542
33 0.73493976 1.000000000           0.2203390
34 0.00000000 0.000000000           0.2033898
35 0.39156627 0.256849315           0.7288136
36 0.34939759 0.363013699           0.6101695
37 0.24698795 0.390410959           0.5932203
38 0.33132530 0.208904110           0.6779661
39 0.15662651 0.441780822           0.9322034
40 0.81927711 0.801369863           0.2711864
41 0.18072289 0.140410959           0.2203390
42 0.74698795 0.489726027           0.4576271
43 0.71686747 0.534246575           0.8135593
44 0.14457831 0.256849315           0.8135593
45 0.08433735 0.010273973           0.0000000
46 0.46385542 0.380136986           0.5254237
47 0.19277108 0.342465753           0.6949153
48 0.29518072 0.123287671           0.1186441
49 0.10843373 0.027397260           0.5762712
50 0.36144578 0.397260274           0.4745763
```

Figure 7: After Normalized Data table.

<u>5.Data Reduction:</u> Given the enormous number of rows in this dataset, it could be more logical to round the murder and assault rates per capita to the nearest zero decimal places, depending on the processing and storage resources we have available.

<u>Code;</u>

> dataset $Murder = format(round(dataset$ Murder ,0))
> dataset $Assault = format(round(dataset$ Assault ,0))
> dataset

```
R 4.2.2 · ~/
> dataset $Murder = format(round(dataset$ Murder ,0))
> dataset
               X 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     170                  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                  57
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
27      Nebraska      4     102                  62
28        Nevada     12     252                  81
29 New Hampshire      2      57                  56
30    New Jersey      7     159                  89
31    New Mexico     11     285                  70
32      New York     11     254                  86
33 North Carolina    13     337                  45
34  North Dakota      1      45                  44
35          Ohio      7     120                  75
36      Oklahoma      7     151                  68
37        Oregon      5     159                  67
38  Pennsylvania      6     106                  72
39  Rhode Island      3     174                  87
40 South Carolina    14     279                  48
41  South Dakota      4      86                  45
42     Tennessee     13     188                  59
43         Texas     13     201                  80
44          Utah      3     120                  80
45       Vermont      2      48                  32
46      Virginia      8     156                  63
47    Washington      4     145                  73
48 West Virginia      6      81                  39
49     Wisconsin      3      53                  66
50       Wyoming      7     161                  60
```

Figure 8: After Reducing value size.

6. Data Discretization: Data discretization is a one kind of reducing process.in can be categorical to numerical. In our dataset 'type' column had four categorical values. which I have replaced by 4 numerical values: "small" replace by 0. "medium" replace by 1. "large" replaced by 2 "extra large" replaced by 3.

Code:

>dataset$PopulationLevel=factor(dataset$OrderedFactorPopulation,levels=c('small', 'medium','large','extra large'), labels = c(1,2,3,4))

> dataset

```
> dataset$OrderedFactorPopulation = factor(dataset$OrderedFactorPopulation,levels = c('small', 'medium','large','extra large'), la
bels = c(1,2,3,4))
> dataset
                X Murder Assault Urban.Population... OrderedFactorPopulation
1         Alabama   13.2     236                  58                       2
2          Alaska   10.0     263                  48                       1
3         Arizona    8.1     294                  80                       4
4        Arkansas    8.8     190                  50                       2
5      California    9.0     276                  91                       4
6        Colorado    7.9     204                  78                       4
7     Connecticut    3.3     110                  77                       4
8        Delaware    5.9     238                  72                       4
9         Florida   15.4     335                  80                       4
10        Georgia   17.4      NA                  60                       3
11         Hawaii    5.3      46                  83                       4
12          Idaho    2.6     120                  54                       2
13       Illinois   10.4     249                  83                       4
14        Indiana    7.2     113                  65                       3
15           Iowa    2.2      56                  57                       2
16         Kansas    6.0     115                  66                       3
17       Kentucky    9.7     109                  52                       2
18      Louisiana   15.4     249                  66                       3
19          Maine    2.1      83                  51                       2
20       Maryland   11.3     300                  67                       3
21  Massachusetts    4.4     149                  85                       4
22       Michigan   12.1     255                  74                       4
23      Minnesota    2.7      72                  66                       3
24    Mississippi   16.1     259                  44                       1
25       Missouri    9.0     178                  70                       4
26        Montana    6.0     109                  53                       2
27       Nebraska    4.3     102                  62                       3
28         Nevada   12.2     252                  81                       4
29  New Hampshire    2.1      57                  56                       2
30     New Jersey    7.4     159                  89                       4
31     New Mexico   11.4     285                  70                       4
32       New York   11.1     254                  86                       4
33 North Carolina   13.0     337                  45                       1
34   North Dakota    0.8      45                  44                       1
35           Ohio    7.3     120                  75                       4
36       Oklahoma    6.6     151                  68                       3
37         Oregon    4.9     159                  67                       3
```

Figure 9: After discretization.

Lastly, after completing all the steps finally we have obtained the full cleaned data set.

	X	Murder	Assault	Urban.Population...	PopulationLevel	OrderedFactorPopulation
1	Alabama	13.2	236.0000	58	medium	2
2	Alaska	10.0	263.0000	48	small	1
3	Arizona	8.1	294.0000	80	extra large	4
4	Arkansas	8.8	190.0000	50	medium	2
5	California	9.0	276.0000	91	extra large	4
6	Colorado	7.9	204.0000	78	extra large	4
7	Connecticut	3.3	110.0000	77	extra large	4
8	Delaware	5.9	238.0000	72	extra large	4
9	Florida	15.4	335.0000	80	extra large	4
10	Georgia	17.4	169.9388	60	large	3
11	Hawaii	5.3	46.0000	83	extra large	4
12	Idaho	2.6	120.0000	54	medium	2
13	Illinois	10.4	249.0000	83	extra large	4
14	Indiana	7.2	113.0000	65	large	3
15	Iowa	2.2	56.0000	57	medium	2
16	Kansas	6.0	115.0000	66	large	3
17	Kentucky	9.7	109.0000	52	medium	2
18	Louisiana	15.4	249.0000	66	large	3
19	Maine	2.1	83.0000	51	medium	2
20	Maryland	11.3	300.0000	67	large	3
21	Massachusetts	4.4	149.0000	85	extra large	4
22	Michigan	12.1	255.0000	74	extra large	4
23	Minnesota	2.7	72.0000	66	large	3
24	Mississippi	16.1	259.0000	44	small	1
25	Missouri	9.0	178.0000	70	extra large	4
26	Montana	6.0	109.0000	53	medium	2
27	Nebraska	4.3	102.0000	62	large	3
28	Nevada	12.2	252.0000	81	extra large	4
29	New Hampshire	2.1	57.0000	56	medium	2
30	New Jersey	7.4	159.0000	89	extra large	4
31	New Mexico	11.4	285.0000	70	extra large	4
32	New York	11.1	254.0000	86	extra large	4
33	North Carolina	13.0	337.0000	45	small	1
34	North Dakota	0.8	45.0000	44	small	1
35	Ohio	7.3	120.0000	75	extra large	4
36	Oklahoma	6.6	151.0000	68	large	3
37	Oregon	4.9	159.0000	67	large	3
38	Pennsylvania	6.3	106.0000	72	extra large	4
39	Rhode Island	3.4	174.0000	87	extra large	4

Figure 10: Full Clean Dataset.

Discussion & Conclusion:

We used R language structures and approaches to gradually improve the data during data processing. After all the data pre-processing procedures were successfully used, the data set was made cleaner and better. Nonetheless, not every technique's step required to be employed for this job. We gained knowledge of the sector's pre-processing of data as well as real data. expanding our toolkit with more knowledge. Data pre-processing helps us increase the dataset's correctness. Any values that are incorrect or missing due to human error or issues are removed. The consistency had increased.