Data Science Project Report

Project Overview: This project is about a dataset that contains statistics in arrests per 100,000 residents for assault and murder, in each of the 50 US states, in 1973. To clean this dataset 3 steps have been maintained. The steps are: Smoothing noisy data, Handling missing data and lastly Data Wrangling. We have integrated a new column named as "Type" which is based on the Urban Population variable. For instance, if Urban population variable's value is less than 50%, in the "Type" column it will show "Small", if less than 60% it will show "Medium", if less than 70% it will show "Large" and "Extra Large" for above 70%. We have also maintained other data pre processing steps such as Data Transformation, Data Reduction and Data Discretization.

Project Solution Design: As a solution to complete the project, different methods of data pre-processing, if else condition, loop iterations have been used. In order to solve this project, we have used R programming language which is a very well-known language for data analysis. As IDE, we have used RStudio to run and compile our code. The dataset that was given was a dirty and noisy dataset. We cleaned the dataset using the data cleaning steps that are: Smoothing Noisy Data, Handling Missing Value and Data Munging. We have also used Data Transformation, Data Reduction and Data Integration.

Data Pre-Processing:

To solve this project, all the necessary steps of data preprocessing have been followed.

Cleaning Data: First of all we loaded the dataset in a csv file and in RStudio, we have read that csv file using read_csv(). After that, we performed the first step of cleaning a dirty data that is handling missing values and smoothing noisy data. We have checked for any missing value in the columns. After that, we rounded the values from decimal points for a smoother data.

Data Integration: In the data integration, as instructed we have created a new column named "Type" which is based on the values of the existing column Urban Population. For example: if the value of urban population is less than 50% then in the new "Type" column, the value will be "Small", for less than 60% in the Urban Population the value will be "Medium" in the Type column and for less than 70% and above 70% it will show "Large" and "Extra Large" respectively in the Type column. We have used for loop and if else conditional statement for data integration.

Data Transformation: For data transformation we have used z-score normalization procedure. We have standardized values using scale(). The scale function makes the computation of z-scores easier and efficient.

Data Reduction: We have used the PCA(Principal Component Analysis) method to reduce the dimensionality of the data. Firstly, we dropped the columns that are categorical because PCA doesn't allow non numeric data. And then we stored the new data frame in another variable called "dataset3". We installed the package "tidyverse" and loaded the data. We used head function to reduce the dataset into first 5 rows. After loading the data, we can use the R built-in function prcomp() to calculate the principal components of the dataset. The eigenvectors in R point in the negative direction by default, so we'll multiply by -1 to reverse the signs. We can see that the first principal component (PC1) has high values for Murder, Assault, and Rape, indicating that it describes the most variation in these variables.

We can also see that the second principal component (PC2) has a high value for Urban Population, indicating that this principle component focuses primarily on urban population. The scores for the principal components for each state are saved in results$x. In order to reverse the signs, we will multiply these scores by -1.

Code and Screenshot of Output:

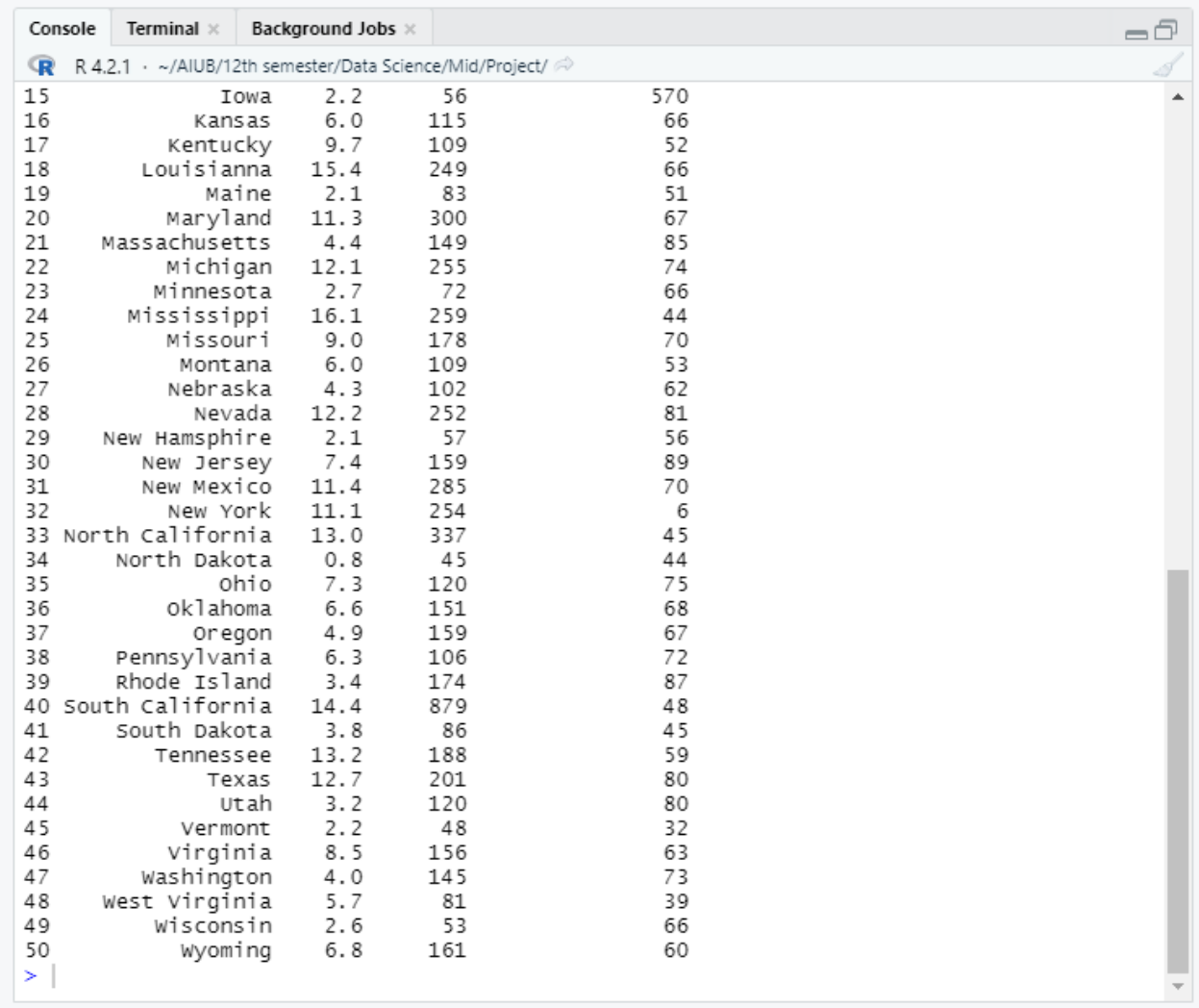
Step 1:

```
print(getwd())

setwd("D:/USER/Documents/AIUB/12th semester/Data Science/Mid/Project")

print(getwd())


dataset<-read.csv("dataset.csv")          #reading csv

print(dataset)
```

Console | Terminal | Background Jobs

R 4.2.1 · ~/AIUB/12th semester/Data Science/Mid/Project/

```
> dataset
                X 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        Ilinois   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      Louisianna   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 Hamsphire    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 California   13.0     337               45
34   North Dakota    0.8      45               44
35           Ohio    7.3     120               75
36       Oklahoma    6.6     151               68
```



15	Iowa	2.2	56	570	
16	Kansas	6.0	115	66	
17	Kentucky	9.7	109	52	
18	Louisianna	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 Hamsphire	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 California	13.0	337	45	
34	North Dakota	0.8	45	44	
35	Ohio	7.3	120	75	
36	Oklahoma	6.6	151	68	
37	Oregon	4.9	159	67	
38	Pennsylvania	6.3	106	72	
39	Rhode Island	3.4	174	87	
40	South California	14.4	879	48	
41	South Dakota	3.8	86	45	
42	Tennessee	13.2	188	59	
43	Texas	12.7	201	80	
44	Utah	3.2	120	80	
45	Vermont	2.2	48	32	
46	Virginia	8.5	156	63	
47	Washington	4.0	145	73	
48	West Virginia	5.7	81	39	
49	Wisconsin	2.6	53	66	
50	Wyoming	6.8	161	60	

Step 2:

#handling missing values

dataset$Murder[is.na(dataset$Murder)]<-mean(dataset$Murder,na.rm=TRUE)

dataset

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

dataset

dataset$Urban_Population[is.na(dataset$Urban_Population)]<-
mean(dataset$Urban_Population,na.rm=TRUE)

dataset

Console | Terminal | Background Jobs

R 4.2.1 · ~/AIUB/12th semester/Data Science/Mid/Project/

```
> #handling missing values
> dataset$Murder[is.na(dataset$Murder)]<-mean(dataset$Murder,na.rm=TRUE)
> 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     182               60
11         Hawaii      5      46               83
12          Idaho      3     120               54
13        Ilinois     10     249               83
14        Indiana      7     113               65
15           Iowa      2      56              570
16         Kansas      6     115               66
17       Kentucky     10     109               52
18     Louisianna     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 Hamsphire      2      57               56
30     New Jersey      7     159               89
31     New Mexico     11     285               70
32       New York     11     254                6
33 North California     13     337               45
```

Console | Terminal | Background Jobs

R 4.2.1 · ~/AIUB/12th semester/Data Science/Mid/Project/

```
33 North California     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 California     14     879               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
```

```
> dataset$Assault[is.na(dataset$Assault)]<-mean(dataset$Assault,na.rm=TRUE)
> 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     182               60
11         Hawaii      5      46               83
12          Idaho      3     120               54
13        Ilinois     10     249               83
14        Indiana      7     113               65
15           Iowa      2      56              570
16         Kansas      6     115               66
17       Kentucky     10     109               52
18      Louisianna     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 Hamsphire      2      57               56
30     New Jersey      7     159               89
31     New Mexico     11     285               70
32       New York     11     254                6
33 North California     13     337               45
34     North Dakota      1      45               44
```

```
33 North California     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 California     14     879              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
> 
```

```
> dataset$Urban_Population[is.na(dataset$Urban_Population)]<-mean(dataset$Urban_Popula
ion,na.rm=TRUE)
> 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     182               60
11         Hawaii      5      46               83
12          Idaho      3     120               54
13        Ilinois     10     249               83
14        Indiana      7     113               65
15           Iowa      2      56              570
16         Kansas      6     115               66
17       Kentucky     10     109               52
18      Louisianna     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 Hamsphire      2      57               56
30     New Jersey      7     159               89
31     New Mexico     11     285               70
30     New Jersey      7     159               89
31     New Mexico     11     285               70
32       New York     11     254                6
33 North California     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 California     14     879               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
>
```

Step 3:

#rounding decimal values

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

dataset

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

dataset

#adding column named 'Type'

dataset2=cbind(dataset,Type=NA)

dataset2

```
> #rounding decimal values
> dataset$Murder =as.numeric(format(round(dataset$Murder, 0)))
> #rounding decimal values
> dataset$Murder =as.numeric(format(round(dataset$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     182               60
11         Hawaii      5      46               83
12          Idaho      3     120               54
13        Ilinois     10     249               83
14        Indiana      7     113               65
15           Iowa      2      56              570
16         Kansas      6     115               66
17       Kentucky     10     109               52
18      Louisianna     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 Hamsphire      2      57               56
30     New Jersey      7     159               89
31     New Mexico     11     285               70
```

```
31       New Mexico     11     285               70
32         New York     11     254                6
33 North California     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 California     14     879               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
> dataset$Assault =as.numeric(format(round(dataset$Assault, 0)))
```

```
> dataset$Assault =as.numeric(format(round(dataset$Assault, 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     182               60
11         Hawaii      5      46               83
12          Idaho      3     120               54
13        Ilinois     10     249               83
14        Indiana      7     113               65
15           Iowa      2      56              570
16         Kansas      6     115               66
17       Kentucky     10     109               52
18     Louisianna     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 Hamsphire      2      57               56
30     New Jersey      7     159               89
31     New Mexico     11     285               70
32       New York     11     254                6
33 North California     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 California     14     879               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
```

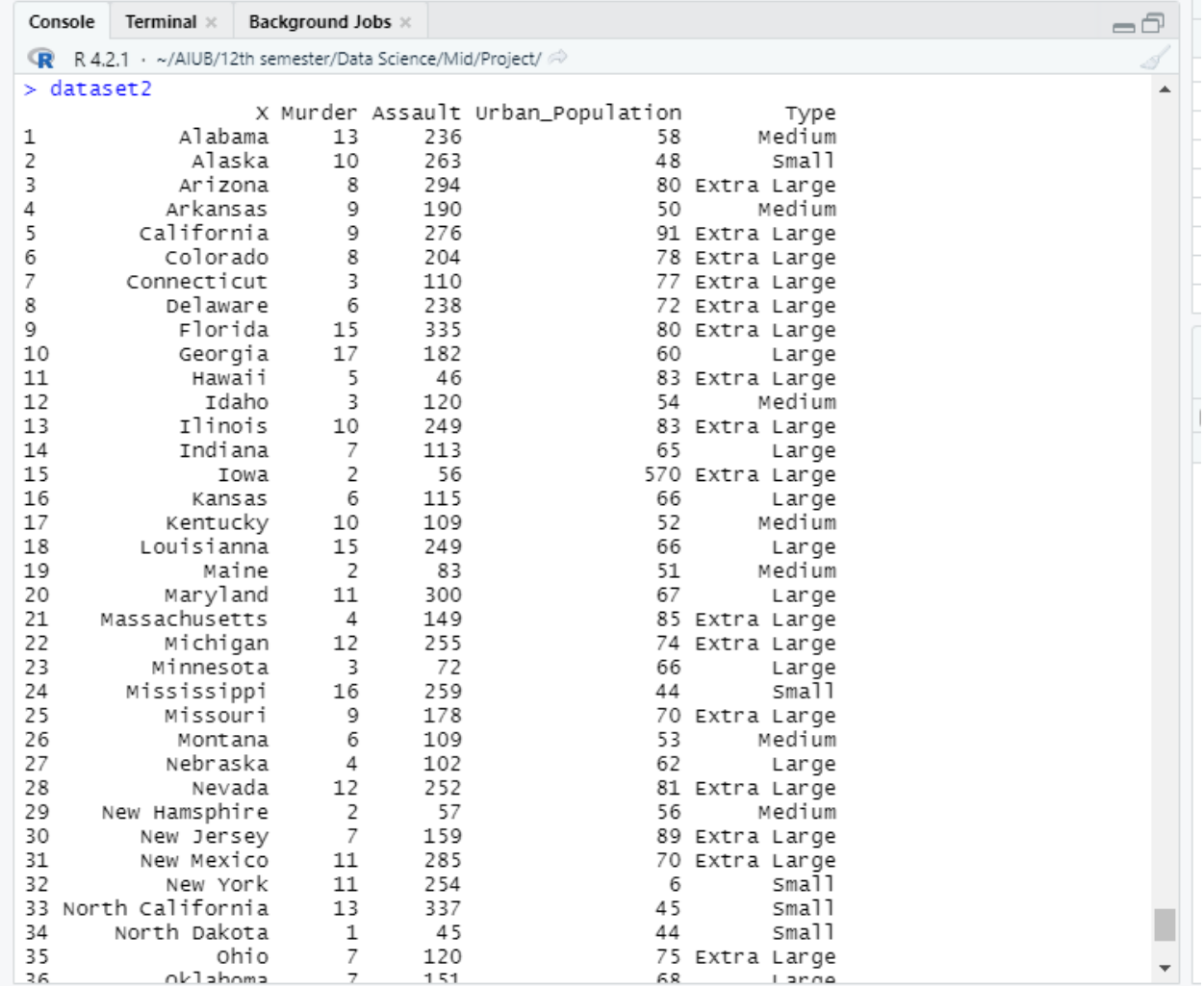
```
> #adding column named 'Type'
> dataset2=cbind(dataset,Type=NA)
> dataset2
                X Murder Assault Urban_Population Type
1         Alabama     13     236                58   NA
2          Alaska     10     263                48   NA
3         Arizona      8     294                80   NA
4        Arkansas      9     190                50   NA
5      California      9     276                91   NA
6        Colorado      8     204                78   NA
7     Connecticut      3     110                77   NA
8        Delaware      6     238                72   NA
9         Florida     15     335                80   NA
10        Georgia     17     182                60   NA
11         Hawaii      5      46                83   NA
12          Idaho      3     120                54   NA
13        Ilinois     10     249                83   NA
14        Indiana      7     113                65   NA
15           Iowa      2      56               570   NA
16         Kansas      6     115                66   NA
17       Kentucky     10     109                52   NA
18     Louisianna     15     249                66   NA
19          Maine      2      83                51   NA
20       Maryland     11     300                67   NA
21  Massachusetts      4     149                85   NA
22       Michigan     12     255                74   NA
23      Minnesota      3      72                66   NA
24    Mississippi     16     259                44   NA
25       Missouri      9     178                70   NA
26        Montana      6     109                53   NA
27       Nebraska      4     102                62   NA
28         Nevada     12     252                81   NA
29  New Hamsphire      2      57                56   NA
30     New Jersey      7     159                89   NA
31     New Mexico     11     285                70   NA
32       New York     11     254                 6   NA
33 North California     13     337                45   NA
34   North Dakota      1      45                44   NA
35           Ohio      7     120                75   NA
36       Oklahoma      7     151                68   NA
37         Oregon      5     159                67   NA
38   Pennsylvania      6     106                72   NA
39   Rhode Island      3     174                87   NA
40 South California     14     879                48   NA
41   South Dakota      4      86                45   NA
42      Tennessee     13     188                59   NA
43          Texas     13     201                80   NA
44           Utah      3     120                80   NA
45        Vermont      2      48                32   NA
46       Virginia      8     156                63   NA
47     Washington      4     145                73   NA
48  West Virginia      6      81                39   NA
49      Wisconsin      3      53                66   NA
50        Wyoming      7     161                60   NA
>
```

Step 4:

#data integration

for(i in 1:nrow(dataset2)) {

```
  if((dataset2$Urban_Population[i])<50){
    dataset2$Type[i]="Small"
  }else if((dataset2$Urban_Population[i])<60){
    dataset2$Type[i]="Medium"

  }else if((dataset2$Urban_Population[i])<70){
    dataset2$Type[i]="Large"

  }else{
    dataset2$Type[i]="Extra Large"
}
  dataset2
```



```
Console   Terminal ×   Background Jobs ×
R 4.2.1 · ~/AIUB/12th semester/Data Science/Mid/Project/
> dataset2
                X Murder Assault Urban_Population        Type
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         Ilinois     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      Louisianna     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 Hamsphire      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 California     13     337               45       Small
34    North Dakota      1      45               44       Small
35            Ohio      7     120               75 Extra Large
36        Oklahoma      7     151               68       Large
37          Oregon      5     159               67       Large
38    Pennsylvania      6     106               72 Extra Large
39    Rhode Island      3     174               87 Extra Large
40 South California     14     879               48       Small
41    South Dakota      4      86               45       Small
42       Tennessee     13     188               59      Medium
43           Texas     13     201               80 Extra Large
44            Utah      3     120               80 Extra Large
45         Vermont      2      48               32       Small
46        Virginia      8     156               63       Large
47      Washington      4     145               73 Extra Large
48   West Virginia      6      81               39       Small
49       Wisconsin      3      53               66       Large
50         Wyoming      7     161               60       Large
>
```

Step 5:

#data transformation using z-score normalization

```
dataset2$Murder_stand2a <- scale(dataset2$Murder)                    # Standardize
using scale()

dataset2$Murder_stand2a


dataset2$Assault_stand2a <- scale(dataset2$Assault)                  # Standardize
using scale()

dataset2$Assault_stand2a


dataset2$Urban_Population_stand2a <- scale(dataset2$Urban_Population)
# Standardize using scale()

dataset2$Urban_Population_stand2a


dataset2
```

Console | Terminal | Background Jobs

R 4.2.1 · ~/AIUB/12th semester/Data Science/Mid/Project/

```
attr(,"scaled:scale")
[1] 73.40828
> dataset2
                X Murder Assault Urban_Population        Type Murder_stand2a
1         Alabama     13     236                58      Medium     1.23769851
2          Alaska     10     263                48       Small     0.53711445
3         Arizona      8     294                80 Extra Large     0.07005841
4        Arkansas      9     190                50      Medium     0.30358643
5      California      9     276                91 Extra Large     0.30358643
6        Colorado      8     204                78 Extra Large     0.07005841
7     Connecticut      3     110                77 Extra Large    -1.09758170
8        Delaware      6     238                72 Extra Large    -0.39699764
9         Florida     15     335                80 Extra Large     1.70475455
10        Georgia     17     182                60       Large     2.17181060
11         Hawaii      5      46                83 Extra Large    -0.63052566
12          Idaho      3     120                54      Medium    -1.09758170
13        Ilinois     10     249                83 Extra Large     0.53711445
14        Indiana      7     113                65       Large    -0.16346961
15           Iowa      2      56               570 Extra Large    -1.33110972
16         Kansas      6     115                66       Large    -0.39699764
17        Kentucky     10     109                52      Medium     0.53711445
18      Louisianna     15     249                66       Large     1.70475455
19           Maine      2      83                51      Medium    -1.33110972
20        Maryland     11     300                67       Large     0.77064247
21   Massachusetts      4     149                85 Extra Large    -0.86405368
22        Michigan     12     255                74 Extra Large     1.00417049
23       Minnesota      3      72                66       Large    -1.09758170
24     Mississippi     16     259                44       Small     1.93828258
25        Missouri      9     178                70 Extra Large     0.30358643
26         Montana      6     109                53      Medium    -0.39699764
27        Nebraska      4     102                62       Large    -0.86405368
28          Nevada     12     252                81 Extra Large     1.00417049
29   New Hamsphire      2      57                56      Medium    -1.33110972
30      New Jersey      7     159                89 Extra Large    -0.16346961
31      New Mexico     11     285                70 Extra Large     0.77064247
32        New York     11     254                 6       Small     0.77064247
33 North California     13     337                45       Small     1.23769851
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

Source

Console | Terminal | Background Jobs

R 4.2.1 · ~/AIUB/12th semester/Data Science/Mid/Project/

33	North California	13	337	45	Small	1.23769851	
34	North Dakota	1	45	44	Small	-1.56463774	
35	Ohio	7	120	75	Extra Large	-0.16346961	
36	Oklahoma	7	151	68	Large	-0.16346961	
37	Oregon	5	159	67	Large	-0.63052566	
38	Pennsylvania	6	106	72	Extra Large	-0.39699764	
39	Rhode Island	3	174	87	Extra Large	-1.09758170	
40	South California	14	879	48	Small	1.47122653	
41	South Dakota	4	86	45	Small	-0.86405368	
42	Tennessee	13	188	59	Medium	1.23769851	
43	Texas	13	201	80	Extra Large	1.23769851	
44	Utah	3	120	80	Extra Large	-1.09758170	
45	Vermont	2	48	32	Small	-1.33110972	
46	Virginia	8	156	63	Large	0.07005841	
47	Washington	4	145	73	Extra Large	-0.86405368	
48	West Virginia	6	81	39	Small	-0.39699764	
49	Wisconsin	3	53	66	Large	-1.09758170	
50	Wyoming	7	161	60	Large	-0.16346961	

	Assault_stand2a	Urban_Population_stand2a
1	0.415485938	-0.220683554
2	0.623923700	-0.356907969
3	0.863241130	0.079010161
4	0.060369752	-0.329663086
5	0.724282623	0.228857019
6	0.168448591	0.051765278
7	-0.557223616	0.038142836
8	0.430925773	-0.029969371
9	1.179757732	0.079010161
10	-0.001389585	-0.193438670
11	-1.051298311	0.119877486
12	-0.480024445	-0.275173320
13	0.515844861	0.119877486
14	-0.534063865	-0.125326463
15	-0.974099140	6.754006535
16	-0.518624031	-0.111704021
17	-0.564943534	-0.302418203
18	0.515844861	-0.111704021

Console | Terminal | Background Jobs

R 4.2.1 · ~/AIUB/12th semester/Data Science/Mid/Project/

33	North California	13	337	45	Small	1.23769851	
34	North Dakota	1	45	44	Small	-1.56463774	
35	Ohio	7	120	75	Extra Large	-0.16346961	
36	Oklahoma	7	151	68	Large	-0.16346961	
37	Oregon	5	159	67	Large	-0.63052566	
38	Pennsylvania	6	106	72	Extra Large	-0.39699764	
39	Rhode Island	3	174	87	Extra Large	-1.09758170	
40	South California	14	879	48	Small	1.47122653	
41	South Dakota	4	86	45	Small	-0.86405368	
42	Tennessee	13	188	59	Medium	1.23769851	
43	Texas	13	201	80	Extra Large	1.23769851	
44	Utah	3	120	80	Extra Large	-1.09758170	
45	Vermont	2	48	32	Small	-1.33110972	
46	Virginia	8	156	63	Large	0.07005841	
47	Washington	4	145	73	Extra Large	-0.86405368	
48	West Virginia	6	81	39	Small	-0.39699764	
49	Wisconsin	3	53	66	Large	-1.09758170	
50	Wyoming	7	161	60	Large	-0.16346961	

	Assault_stand2a	Urban_Population_stand2a
1	0.415485938	-0.220683554
2	0.623923700	-0.356907969
3	0.863241130	0.079010161
4	0.060369752	-0.329663086
5	0.724282623	0.228857019
6	0.168448591	0.051765278
7	-0.557223616	0.038142836
8	0.430925773	-0.029969371
9	1.179757732	0.079010161
10	-0.001389585	-0.193438670
11	-1.051298311	0.119877486
12	-0.480024445	-0.275173320
13	0.515844861	0.119877486
14	-0.534063865	-0.125326463
15	-0.974099140	6.754006535
16	-0.518624031	-0.111704021
17	-0.564943534	-0.302418203
18	0.515844861	-0.111704021

```
18     0.515844861             -0.111704021
19    -0.765661378             -0.316040645
20     0.909560633             -0.098081579
21    -0.256146849              0.147122369
22     0.562164363             -0.002724488
23    -0.850580466             -0.111704021
24     0.593044032             -0.411397736
25    -0.032269253             -0.057214255
26    -0.564943534             -0.288795761
27    -0.618982953             -0.166193787
28     0.539004612              0.092632603
29    -0.966379223             -0.247928437
30    -0.178947678              0.201612135
31     0.793761876             -0.057214255
32     0.554444446             -0.929050516
33     1.195197566             -0.397775294
34    -1.059018228             -0.411397736
35    -0.480024445              0.010897953
36    -0.240707015             -0.084459138
37    -0.178947678             -0.098081579
38    -0.588103285             -0.029969371
39    -0.063148922              0.174367252
40     5.379392635             -0.356907969
41    -0.742501627             -0.397775294
42     0.044929918             -0.207061112
43     0.145288840              0.079010161
44    -0.480024445              0.079010161
45    -1.035858477             -0.574867035
46    -0.202107430             -0.152571346
47    -0.287026518             -0.016346930
48    -0.781101212             -0.479509944
49    -0.997258891             -0.111704021
50    -0.163507844             -0.193438670
> |
```

Step 6:

```
#data Reduction using PCA

library(tidyverse)

dataset3 <- dataset2[,-c(1,5)]

dataset3

head(dataset3)

results <- prcomp(dataset3 , scale = TRUE)

results$rotation <- -1*results$rotation

results$rotation

results$x <- -1*results$x

head(results$x)

head(dataset3[order(-dataset3$Murder),])

results$sdev^2 / sum(results$sdev^2)
```

Console | Terminal | Background Jobs

R 4.2.1 · ~/AIUB/12th semester/Data Science/Mid/Project/

```
49              -0.111704021
50              -0.193438670
> head(dataset3)
  Murder Assault Urban_Population Murder_stand2a Assault_stand2a
1     13     236               58     1.23769851      0.41548594
2     10     263               48     0.53711445      0.62392370
3      8     294               80     0.07005841      0.86324113
4      9     190               50     0.30358643      0.06036975
5      9     276               91     0.30358643      0.72428262
6      8     204               78     0.07005841      0.16844859
  Urban_Population_stand2a
1               -0.22068355
2               -0.35690797
3                0.07901016
4               -0.32966309
5                0.22885702
6                0.05176528
> results <- prcomp(dataset3 , scale = TRUE)
> results$rotation <- -1*results$rotation
> results$rotation
                                PC1        PC2          PC3          PC4          PC5
Murder                    0.4792175 -0.1238830 -0.50497876  0.02232623 -0.00542836
Assault                   0.4700700 -0.1902892  0.49277198  0.27918815 -0.64951004
Urban_Population         -0.2222268 -0.6696589 -0.04660715  0.64927307  0.27947672
Murder_stand2a            0.4792175 -0.1238830 -0.50497876 -0.02232623  0.00542836
Assault_stand2a           0.4700700 -0.1902892  0.49277198 -0.27918815  0.64951004
Urban_Population_stand2a -0.2222268 -0.6696589 -0.04660715 -0.64927307 -0.27947672
                                PC6
Murder                    0.70673338
Assault                  -0.01380859
Urban_Population         -0.01836438
Murder_stand2a           -0.70673338
Assault_stand2a           0.01380859
Urban_Population_stand2a  0.01836438
> results$x <- -1*results$x
> head(results$x)
          PC1        PC2         PC3            PC4            PC5            PC6
[1,] 1.6749522 -0.1692192 -0.81997239  5.551115e-17 -1.595946e-16 -6.808790e-16
```

```
Console   Terminal ×   Background Jobs ×
R 4.2.1 · ~/AIUB/12th semester/Data Science/Mid/Project/
6               0.05176528
> results <- prcomp(dataset3 , scale = TRUE)
> results$rotation <- -1*results$rotation
> results$rotation
                                PC1        PC2          PC3          PC4          PC5
Murder                    0.4792175 -0.1238830 -0.50497876  0.02232623 -0.00542836
Assault                   0.4700700 -0.1902892  0.49277198  0.27918815 -0.64951004
Urban_Population         -0.2222268 -0.6696589 -0.04660715  0.64927307  0.27947672
Murder_stand2a            0.4792175 -0.1238830 -0.50497876 -0.02232623  0.00542836
Assault_stand2a           0.4700700 -0.1902892  0.49277198 -0.27918815  0.64951004
Urban_Population_stand2a -0.2222268 -0.6696589 -0.04660715 -0.64927307 -0.27947672
                                PC6
Murder                    0.70673338
Assault                  -0.01380859
Urban_Population         -0.01836438
Murder_stand2a           -0.70673338
Assault_stand2a           0.01380859
Urban_Population_stand2a  0.01836438
> results$x <- -1*results$x
> head(results$x)
            PC1        PC2         PC3           PC4           PC5           PC6
[1,] 1.6749522 -0.1692192 -0.81997239  5.551115e-17 -1.595946e-16 -6.808790e-16
[2,] 1.2599940  0.1074826  0.10571039  1.665335e-16 -1.942890e-16 -3.044440e-16
[3,] 0.8435976 -0.4517087  0.77264120  2.775558e-17 -2.706169e-16  6.808790e-17
[4,] 0.4942438  0.3433298 -0.21638303  1.110223e-16  1.110223e-16 -3.061787e-16
[5,] 0.8701786 -0.6573770  0.38587023 -8.326673e-17 -4.302114e-16 -9.107298e-17
[6,] 0.2025044 -0.1507961  0.09043221 -4.857226e-17 -7.459311e-17 -2.699663e-17
> head(dataset3[order(-dataset3$Murder),])
   Murder Assault Urban_Population Murder_stand2a Assault_stand2a
10     17     182               60       2.171811    -0.001389585
24     16     259               44       1.938283     0.593044032
9      15     335               80       1.704755     1.179757732
18     15     249               66       1.704755     0.515844861
40     14     879               48       1.471227     5.379392635
1      13     236               58       1.237699     0.415485938
   Urban_Population_stand2a
10              -0.19343867
24              -0.41139774
```

```
R 4.2.1 · ~/AIUB/12th semester/Data Science/Mid/Project/
Murder_stand2a            0.4792175 -0.1238830 -0.50497876 -0.02232623  0.00542836
Assault_stand2a           0.4700700 -0.1902892  0.49277198 -0.27918815  0.64951004
Urban_Population_stand2a -0.2222268 -0.6696589 -0.04660715 -0.64927307 -0.27947672
                                PC6
Murder                   0.70673338
Assault                 -0.01380859
Urban_Population        -0.01836438
Murder_stand2a          -0.70673338
Assault_stand2a          0.01380859
Urban_Population_stand2a 0.01836438
> results$x <- -1*results$x
> head(results$x)
           PC1        PC2         PC3           PC4           PC5           PC6
[1,] 1.6749522 -0.1692192 -0.81997239  5.551115e-17 -1.595946e-16 -6.808790e-16
[2,] 1.2599940  0.1074826  0.10571039  1.665335e-16 -1.942890e-16 -3.044440e-16
[3,] 0.8435976 -0.4517087  0.77264120  2.775558e-17 -2.706169e-16  6.808790e-17
[4,] 0.4942438  0.3433298 -0.21638303  1.110223e-16  1.110223e-16 -3.061787e-16
[5,] 0.8701786 -0.6573770  0.38587023 -8.326673e-17 -4.302114e-16 -9.107298e-17
[6,] 0.2025044 -0.1507961  0.09043221 -4.857226e-17 -7.459311e-17 -2.699663e-17
> head(dataset3[order(-dataset3$Murder),])
   Murder Assault Urban_Population Murder_stand2a Assault_stand2a
10     17     182               60       2.171811    -0.001389585
24     16     259               44       1.938283     0.593044032
9      15     335               80       1.704755     1.179757732
18     15     249               66       1.704755     0.515844861
40     14     879               48       1.471227     5.379392635
1      13     236               58       1.237699     0.415485938
   Urban_Population_stand2a
10              -0.19343867
24              -0.41139774
9                0.07901016
18              -0.11170402
40              -0.35690797
1               -0.22068355
> results$sdev^2 / sum(results$sdev^2)
[1] 5.740050e-01 3.078718e-01 1.181231e-01 2.691532e-32 1.713257e-32 2.446159e-33
>
```

Discussion and Conclusion: We attempted to process the given dataset in such a way that it is clean, easy to analyze, and manage. We completed our project by following all of the steps required for a complete and clean dataset. This dataset can now be used to analyze data very conveniently.