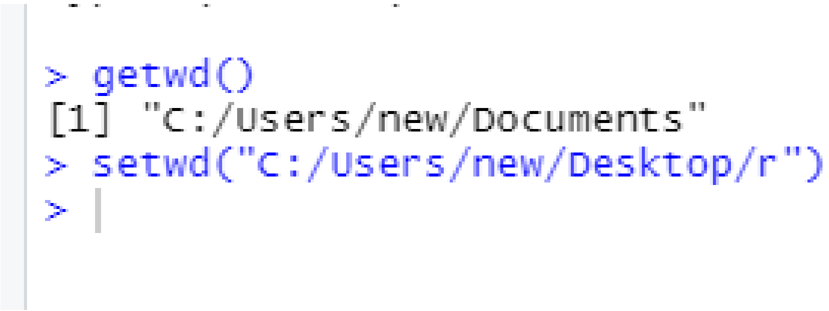
Project Title: Apply Data Pre-processing on a Dataset

Project Overview:

The following dataset 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.

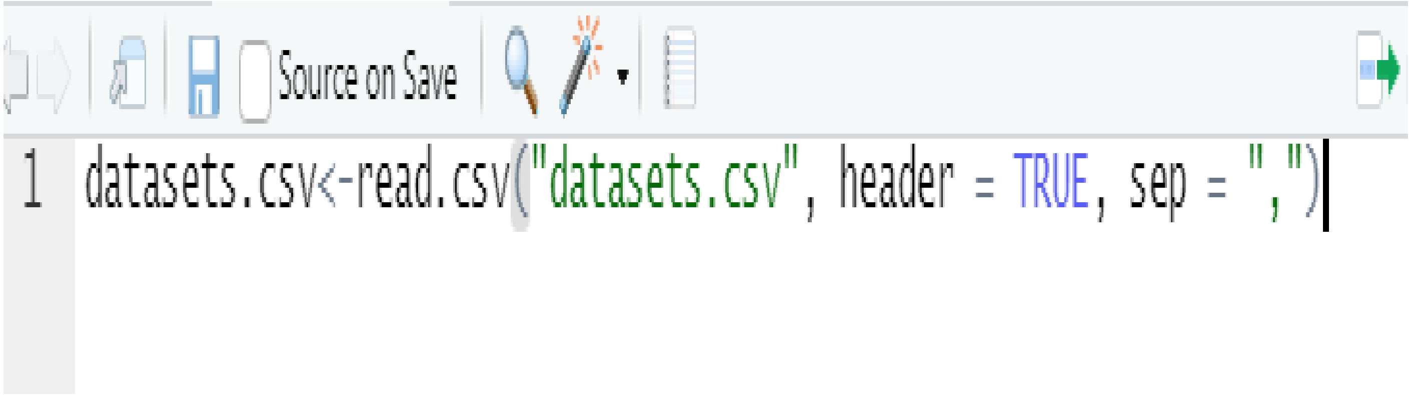
1.Data cleaning :

Import dataset from excel to R-studio: At first I have to set the working directory. By using setwd(“My Location”) command.



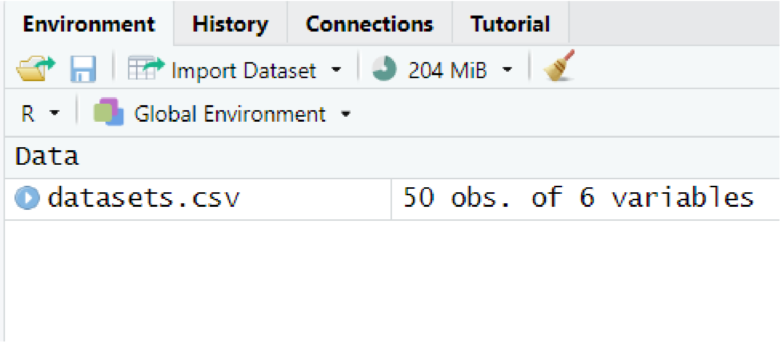
```
> getwd()
[1] "C:/Users/new/Documents"
> setwd("C:/Users/new/Desktop/r")
> 
```

Here, I can set my directory where I can work. After set directory now i should insert table from excel file to R-STUDIO using R-language code.



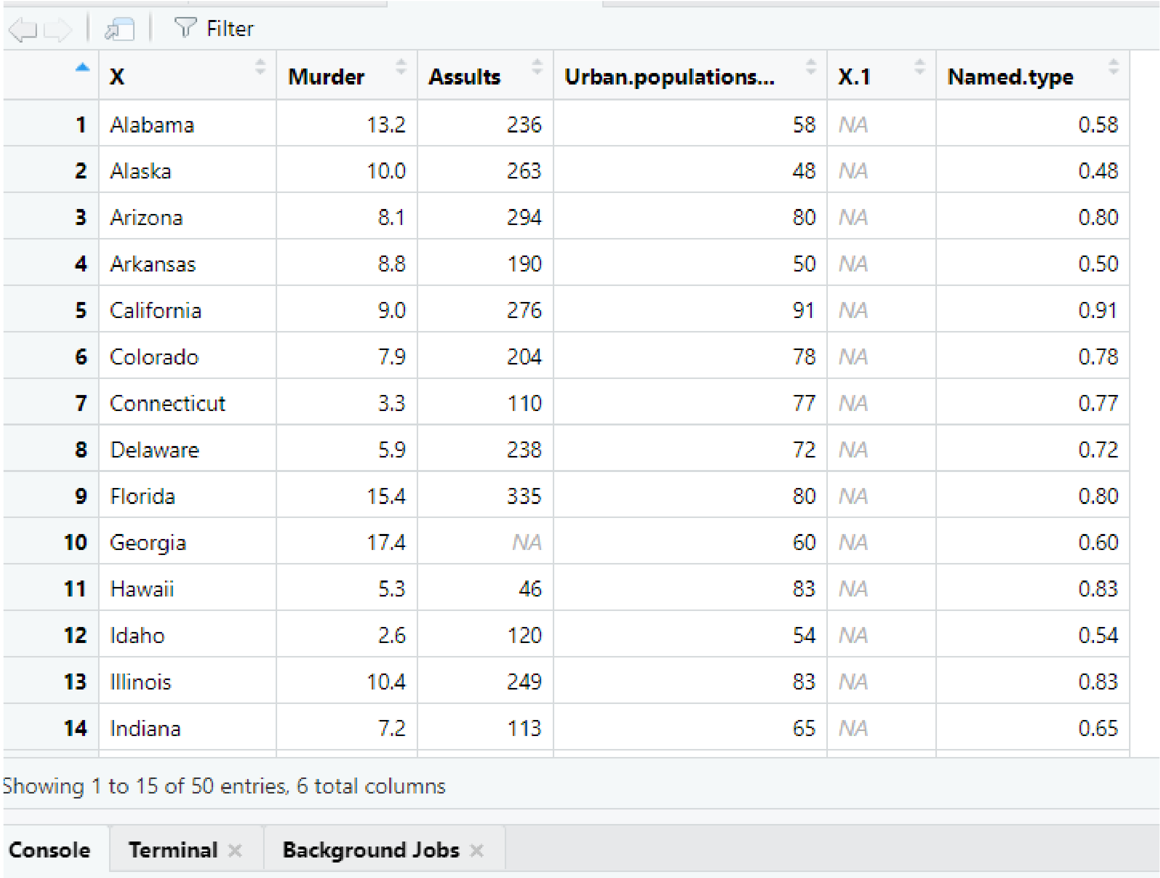
```
datasets.csv<-read.csv("datasets.csv", header = TRUE, sep = ",")
```

If you can see that datasets is my excel csv file name and using this code I’m transfer excel dataset to R.



```
Data
datasets.csv          50 obs. of 6 variables
```

Here is my imported table:

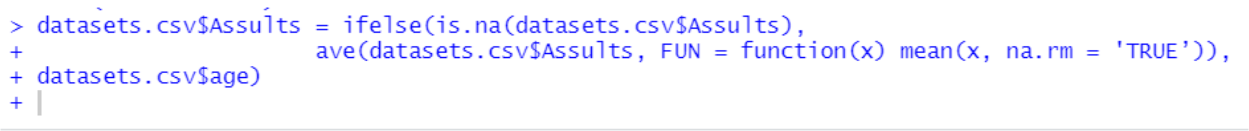


	X	Murder	Assults	Urban.populations...	X.1	Named.type
1	Alabama	13.2	236	58	NA	0.58
2	Alaska	10.0	263	48	NA	0.48
3	Arizona	8.1	294	80	NA	0.80
4	Arkansas	8.8	190	50	NA	0.50
5	California	9.0	276	91	NA	0.91
6	Colorado	7.9	204	78	NA	0.78
7	Connecticut	3.3	110	77	NA	0.77
8	Delaware	5.9	238	72	NA	0.72
9	Florida	15.4	335	80	NA	0.80
10	Georgia	17.4	NA	60	NA	0.60
11	Hawaii	5.3	46	83	NA	0.83
12	Idaho	2.6	120	54	NA	0.54
13	Illinois	10.4	249	83	NA	0.83
14	Indiana	7.2	113	65	NA	0.65

Showing 1 to 15 of 50 entries, 6 total columns

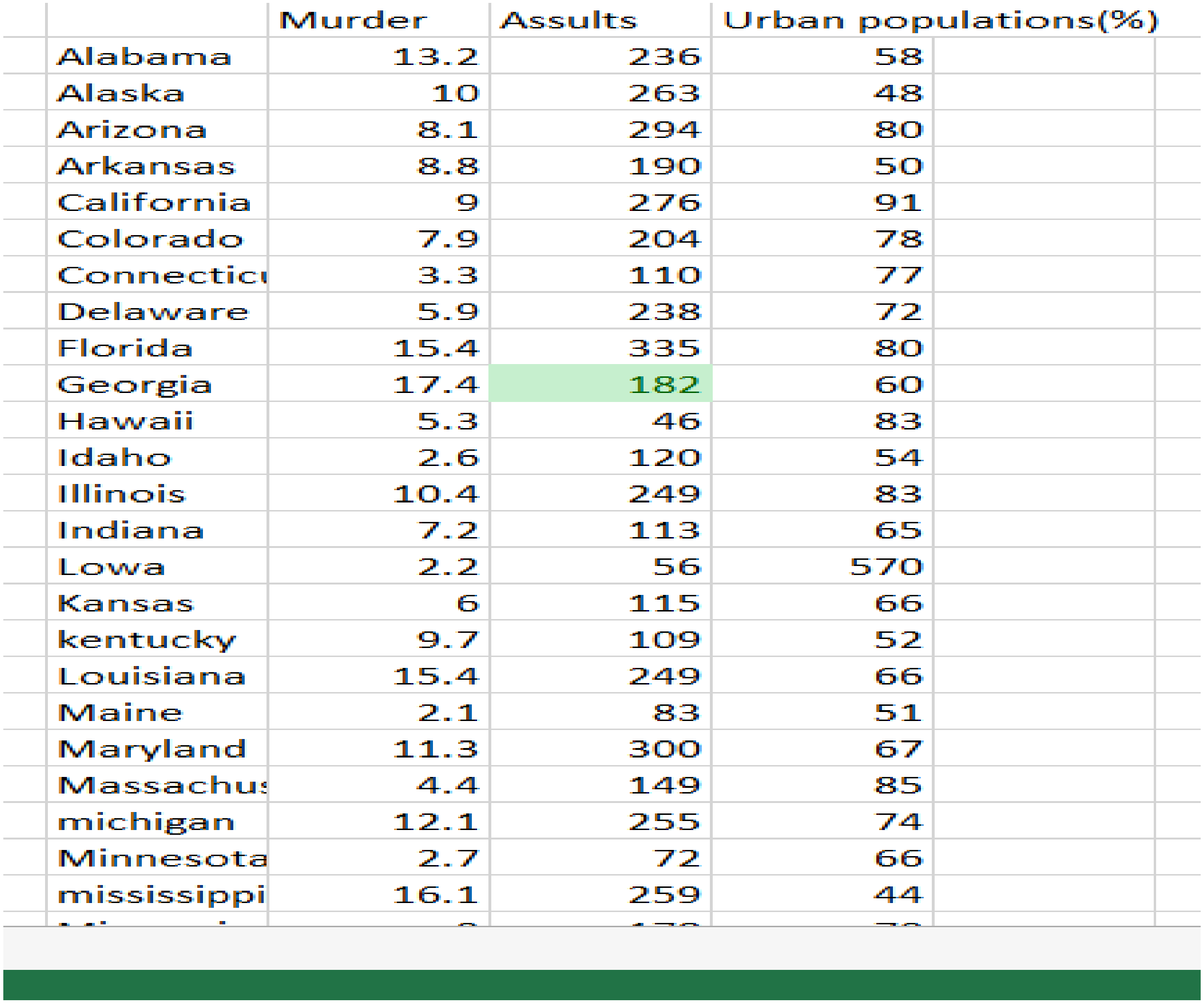
Next step-(Dealing with missing values):

Here we can see that in the table inside the Assults column, in number ten value is missing (NA). Now, we can replace the missing values by the mean values of the respective variables. By using this code I can find my missing value. Inside the code datasets is my csv file name. Inside ifelse conditions Assults is my table column name. FUN is a function of x that calculates the mean excluding NA values (na.rm=TRUE).



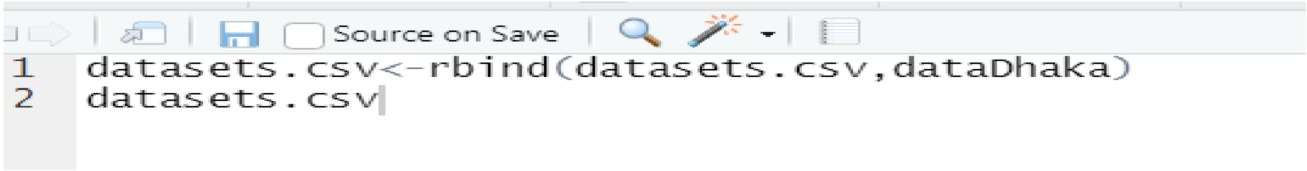
```
> datasets.csv$Assults = ifelse(is.na(datasets.csv$Assults),
+                         ave(datasets.csv$Assults, FUN = function(x) mean(x, na.rm = 'TRUE')),
+ datasets.csv$age)
+ |
```

I find missing value by using sum of assults column value then divided into total state(49).After that I find missing value(182).



	Murder	Assults	Urban populations(%)
Alabama	13.2	236	58
Alaska	10	263	48
Arizona	8.1	294	80
Arkansas	8.8	190	50
California	9	276	91
Colorado	7.9	204	78
Connecticu	3.3	110	77
Delaware	5.9	238	72
Florida	15.4	335	80
Georgia	17.4	182	60
Hawaii	5.3	46	83
Idaho	2.6	120	54
Illinois	10.4	249	83
Indiana	7.2	113	65
Lowa	2.2	56	570
Kansas	6	115	66
kentucky	9.7	109	52
Louisiana	15.4	249	66
Maine	2.1	83	51
Maryland	11.3	300	67
Massachus	4.4	149	85
michigan	12.1	255	74
Minnesota	2.7	72	66
mississippi	16.1	259	44

2. Data integration:



```
datasets.csv<-rbind(datasets.csv,dataDhaka)
datasets.csv
```

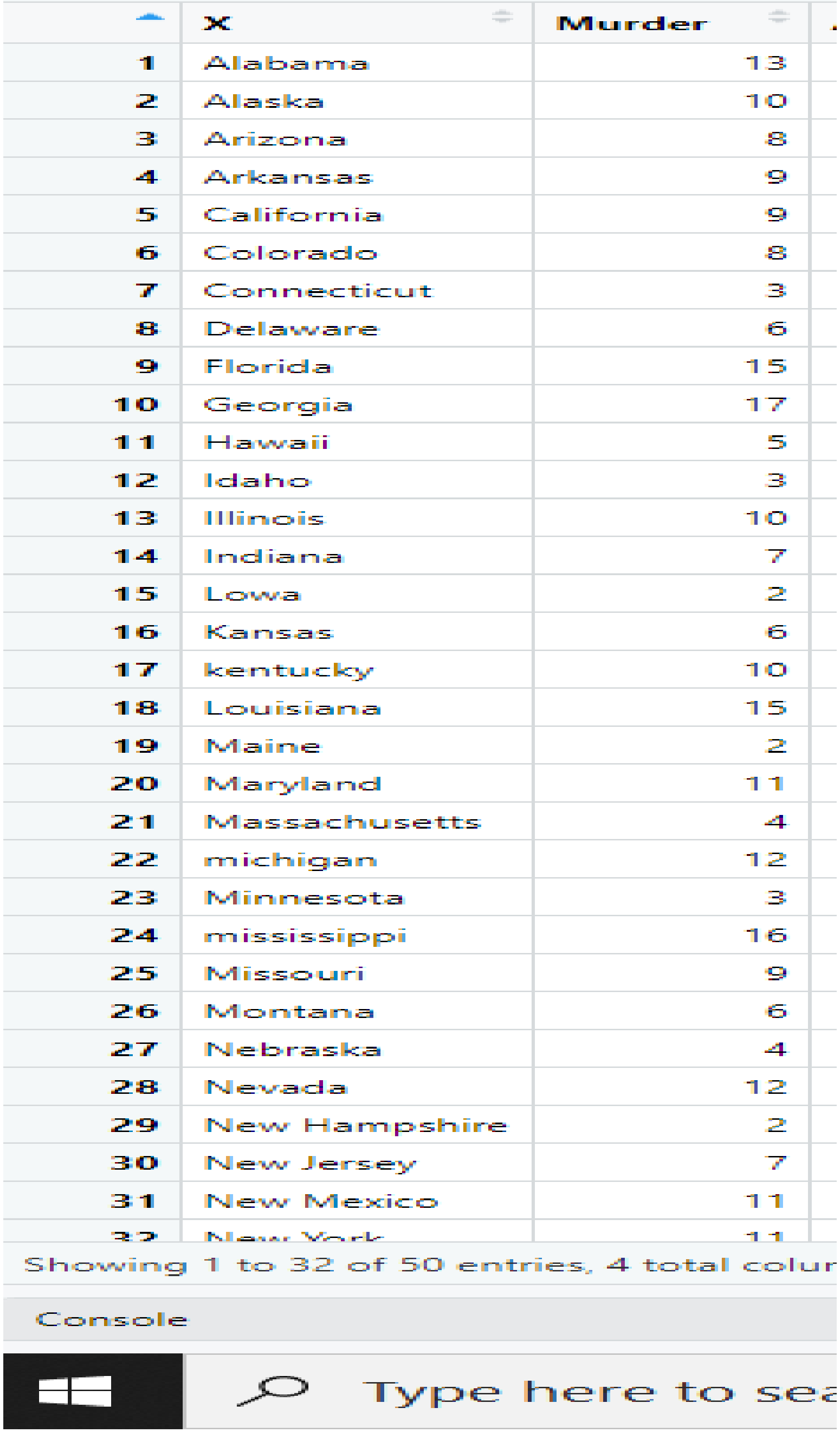
By using this code I can combine data from different source. We can easily integrate data by using rbind function.If i want to add Dhaka dataset then i use this function for integrate data.

3.Data Transformation:

In the dataset, we can see that the Murder variable in the dataset contains decimal value in the data .But I’m not interested in having decimal places for Murder variable, i can round it up.

Code :

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

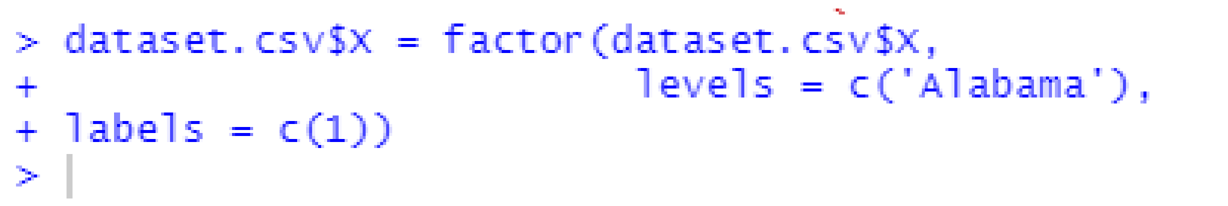


	X	Murder
1	Alabama	13
2	Alaska	10
3	Arizona	8
4	Arkansas	9
5	California	9
6	Colorado	8
7	Connecticut	3
8	Delaware	6
9	Florida	15
10	Georgia	17
11	Hawaii	5
12	Idaho	3
13	Illinois	10
14	Indiana	7
15	Lowa	2
16	Kansas	6
17	kentucky	10
18	Louisiana	15
19	Maine	2
20	Maryland	11
21	Massachusetts	4
22	michigan	12
23	Minnesota	3
24	mississippi	16
25	Missouri	9
26	Montana	6
27	Nebraska	4
28	Nevada	12
29	New Hampshire	2
30	New Jersey	7
31	New Mexico	11
32	New York	11

Showing 1 to 32 of 50 entries, 4 total colur

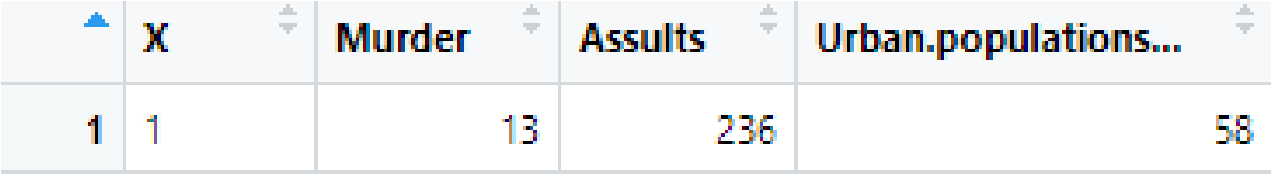
By using data transformations formula, decimal values are gone from the dataset.

Categorical Data:



```
> dataset.csv$X = factor(dataset.csv$X,
+                         levels = c('Alabama'),
+ labels = c(1))
> |
```

By using this code I convert the categorical features as numerical codes.



	X	Murder	Assults	Urban.populations...
1	1	13	236	58

Here i change Alabama state name as state code. Now code 1 is for Alabama state. It’s help to recognize state very easily.

4.Data Reduction:

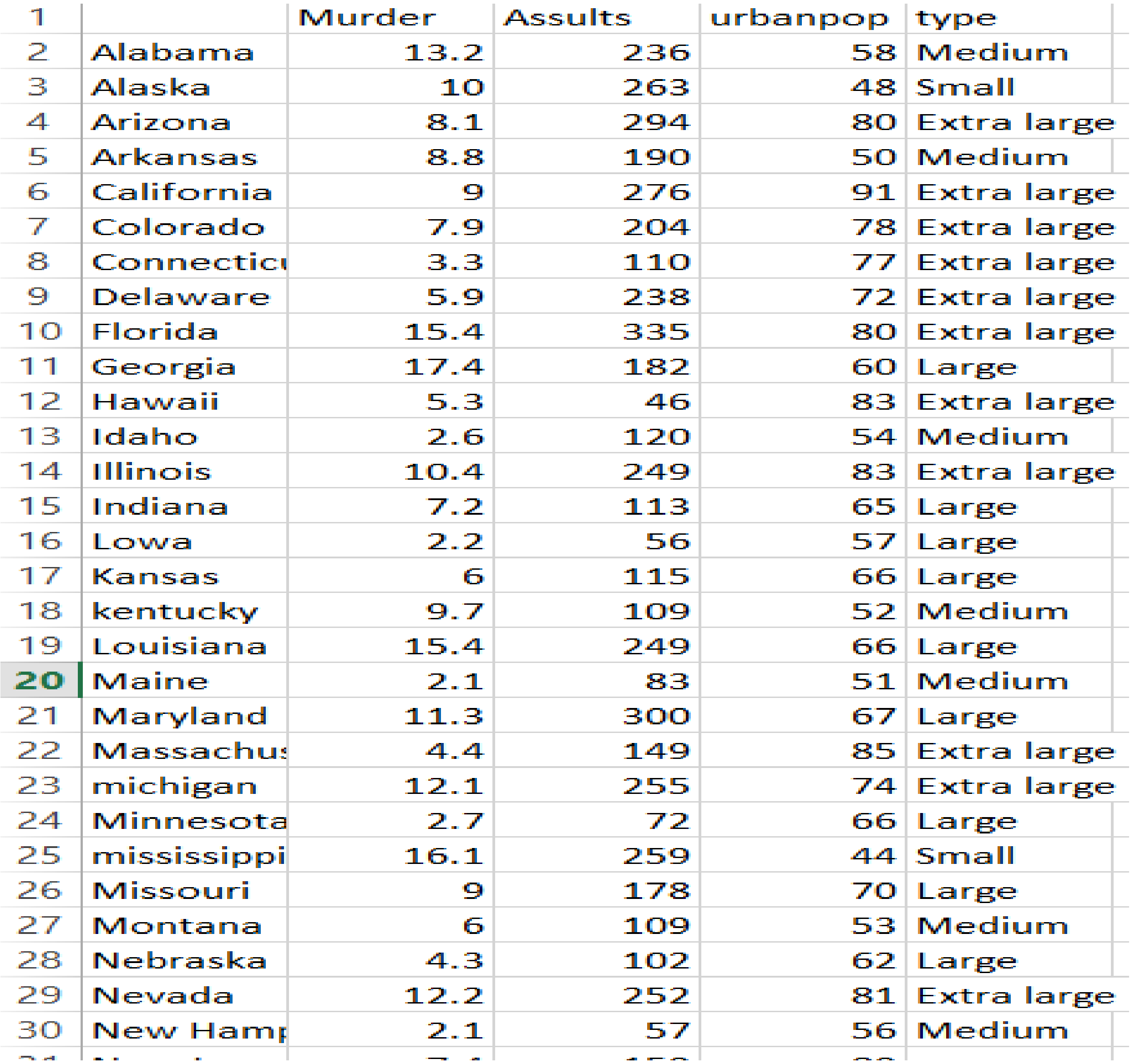
Data reduction is the process of reducing the amount of capacity required to store data by using this we can reduce data without changing information.

5.Data Discretization:

Is a process to convert data in categorical using this code we can make it more easier or make it grouping. In the ifelse condition I implement that <50 urban population is small ,same as <60 is medium and <70 is large .

Code:

```
datasetss.csv<-
transform(datasetss.csv,type=ifelse(Urbanpop<50,'Small',ifelse
(Urbanpop<60,'Medium',ifelse(Urbanpop<70,'Large',"Extra
large"))))
```



1		Murder	Assults	urbanpop	type
2	Alabama	13.2	236	58	Medium
3	Alaska	10	263	48	Small
4	Arizona	8.1	294	80	Extra large
5	Arkansas	8.8	190	50	Medium
6	California	9	276	91	Extra large
7	Colorado	7.9	204	78	Extra large
8	Connecticu	3.3	110	77	Extra large
9	Delaware	5.9	238	72	Extra large
10	Florida	15.4	335	80	Extra large
11	Georgia	17.4	182	60	Large
12	Hawaii	5.3	46	83	Extra large
13	Idaho	2.6	120	54	Medium
14	Illinois	10.4	249	83	Extra large
15	Indiana	7.2	113	65	Large
16	Lowa	2.2	56	57	Large
17	Kansas	6	115	66	Large
18	kentucky	9.7	109	52	Medium
19	Louisiana	15.4	249	66	Large
20	Maine	2.1	83	51	Medium
21	Maryland	11.3	300	67	Large
22	Massachus	4.4	149	85	Extra large
23	michigan	12.1	255	74	Extra large
24	Minnesota	2.7	72	66	Large
25	mississippi	16.1	259	44	Small
26	Missouri	9	178	70	Large
27	Montana	6	109	53	Medium
28	Nebraska	4.3	102	62	Large
29	Nevada	12.2	252	81	Extra large
30	New Hamp	2.1	57	56	Medium

Discussion and Conclusion:

After doing the data pre- processing,data cleaning , Smooth Noisy Data, Handling Missing Data, Data Wrangling we find a clean dataset and by removing NA from dataset. NA are replace with average value.In data transformation process from dataset, i can removing decimal places for Murder variable, i can round it up from data transformations I find a good dataset and it’s easy to human understandable value not float value. In data discretization i can convert data in categorical from type column now we can show data as categorically .Form completed all steps we can find a perfect dataset and datatable.