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.

We have to apply the pre-processing techniques to prepare the dataset for data analysis. In order to prepare a cleaned dataset, we need to perform the following tasks of data pre-processing using R language:

1. Data cleaning:

- Smooth Noisy Data
- Handling Missing Data
- Data Wrangling or Munging

2. Data Integration

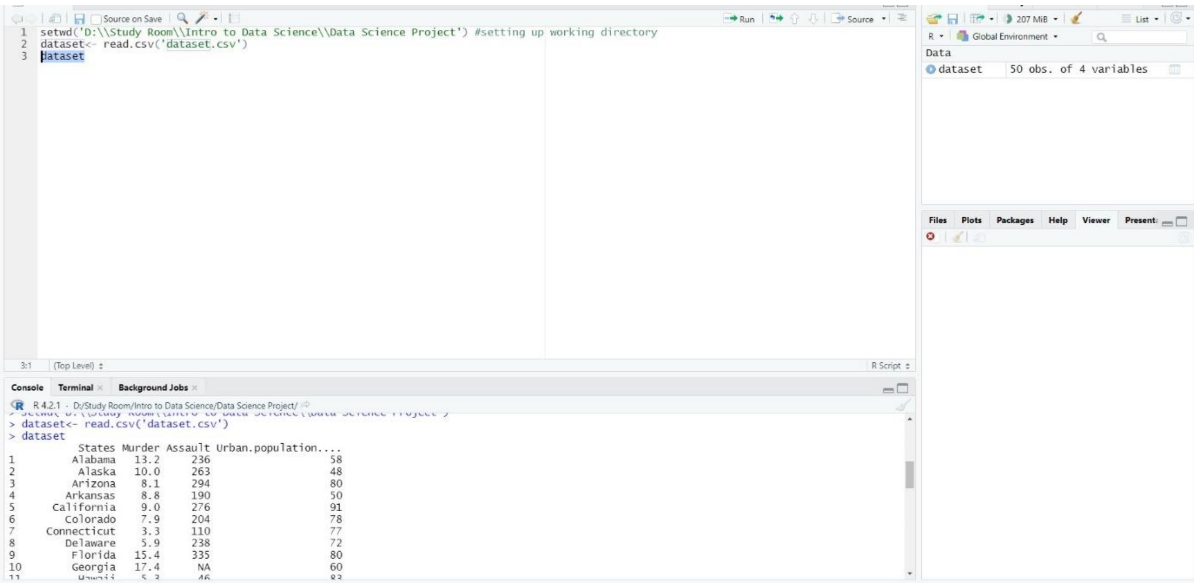
3. Data Transformation

4. Data Reduction

5. Data Discretization

Project Solution Design:

Firstly, the data file was scanned and transformed into a csv(comma separated value) format. Then, Rstudio tool was used and the csv file was read and imported as a dataframe in Rstudio. Next, the data cleaning process started, first missing values were checked and we handled the missing values. Then we started to handle noisy data, we checked the outliers for different variables and handle the outliers with proper methods. Then we started data integration part and add a new column in the dataframe which contains value according to the given condition comparing other variable of the particular row. Then we performed data transformation where the values are converted decimal values in integer numbers. Lastly, data discretization was performed to categorized and understand the data easily.

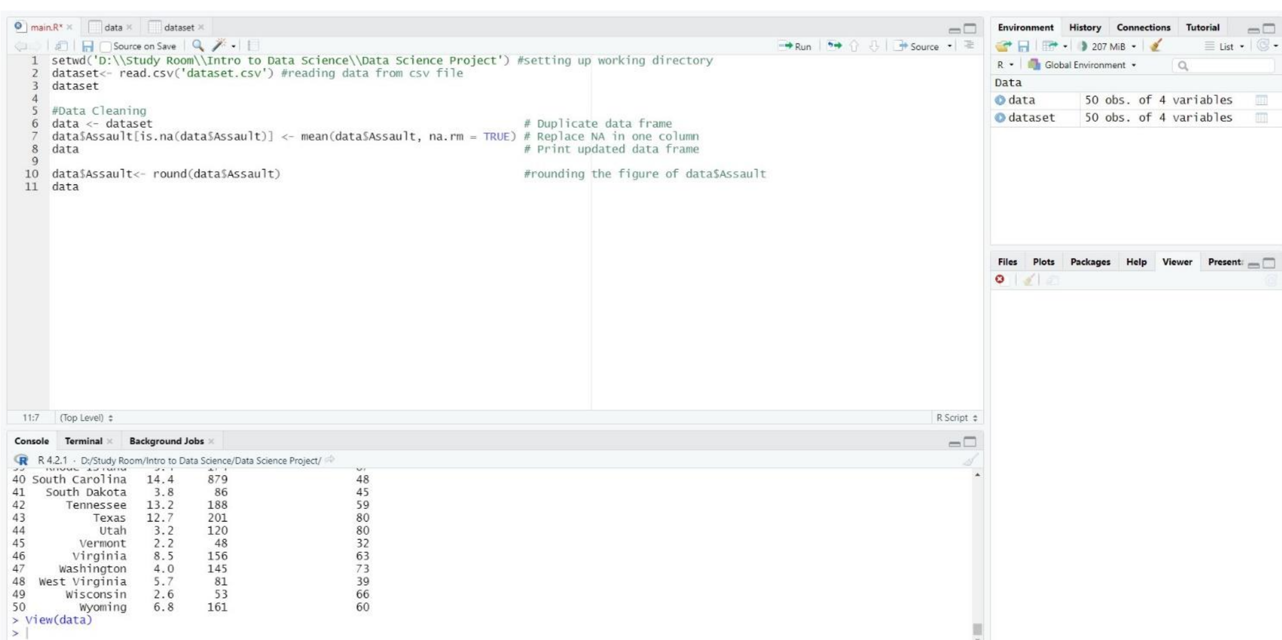


1. Data Cleaning:

Handling Missing Data – In order to handle missing data at first we check if there are any missing values in the dataframe.

```
> any(is.na(dataset))
[1] TRUE
```

The result shows true, that means there are missing values in the data. And then we found the there were missing value in the "Assault" column. In order to handle the missing value we replace it with the Mean value of that column.



Though it is a part of data transformation, here, we have also round the values of the column "Assault" as the mean value was four decimal numbers, all the numbers converted into four decimal values. So, we used round function to fix it.

Before Handling Missing Values: (50 rows, 4variables)

	States	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
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
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 Carolina	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

After Handling Missing Values: (49 rows, 4 variables)

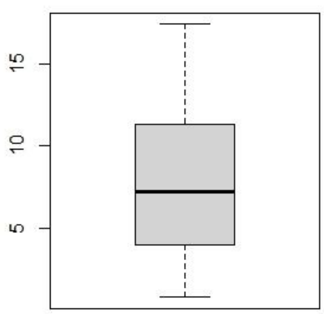
	States	Murder	Assault	Urban.population....
1	Alabama	13.2	236	58
2	Alaska	10.0	263	48
3	Arizona	8.1	294	80
4	Arkansas	8.8	190	50
5	California	9.0	276	91
6	Colorado	7.9	204	78
7	Connecticut	3.3	110	77
8	Delaware	5.9	238	72
9	Florida	15.4	335	80
10	Georgia	17.4	182	60
11	Hawaii	5.3	46	83
12	Idaho	2.6	120	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
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
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 Carolina	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

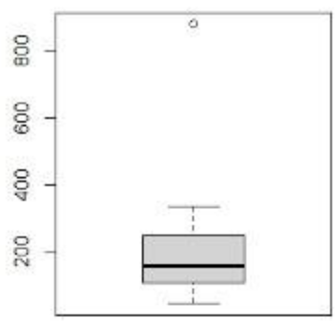
Smoothing Noisy Data:

In order to smooth noisy data, we need to find outliers in different variables and then we will remove the noisy data. In order to find the outlier we used boxplot() functions, it helps to detect outliers easily.

```r
setwd('D:\\Study Room\\Intro to Data Science\\Data Science Project') #setting up working directory
dataset<- read.csv('dataset.csv') #reading data from csv file
dataset

#Data Cleaning
data <- dataset                                                    # Duplicate data frame
data$Assault[is.na(data$Assault)] <- mean(data$Assault, na.rm = TRUE) # Replace NA in one column
data                                                               # Print updated data frame

data$Assault<- round(data$Assault)                                 #rounding the figure of data$Assault
data

#Smothing Noisy Data

boxplot(data$Assault)              #finding outliers in the Assault column using box plot
boxplot(data$Urban.population....) #finding outliers in the Urban Population column using box plot
boxplot(data$Murder)               #finding outliers in the Murder column using box plot

```

After plotting:





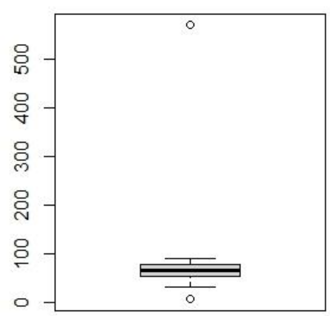


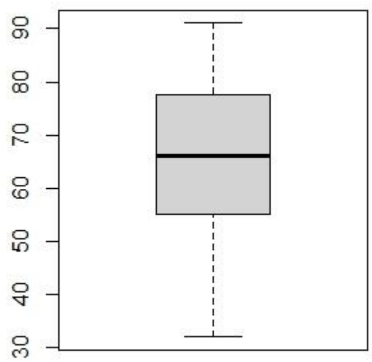
Fig 1- Boxplot for Murder Fig 2- Boxplot for Assault Fig 3- Boxplot for Population

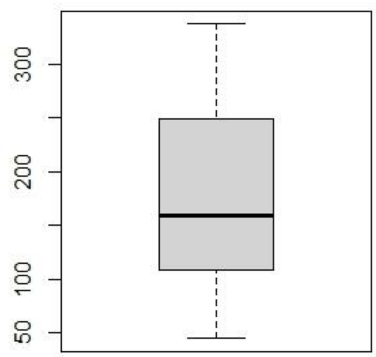
From here, we can see there is no outlier in Murder column but in the Assault column there is one outlier and in the Population column there is 2 outliers. So, we have to handle these noisy data by taking some thresholds for the value and using condition to check the values of each column whether they are between the thresholds value.

For population, as it is in percentage, we kept the threshold value between 10 to 100, any value greater than 100 or less than 10 was removed. So, rows that contains population percentage more than 100 or less than 10 was dropped out. For assaults, we kept the threshold values less than 400.

```r

#removing outliers
data2 <- data[data$Assault<400, ]     # Remove outliers for Assault column where we select the threshold value 400
boxplot(data2$Assault)                #checking if there is more outliers

data3<- data2[(data$Urban.population....<101) & (data$Urban.population....>10), ] # Remove outliers for Assault column
boxplot(data3$Urban.population....)   #checking if there is more outliers


```

Then we again checked if there were any outliers remain by using boxplot()A





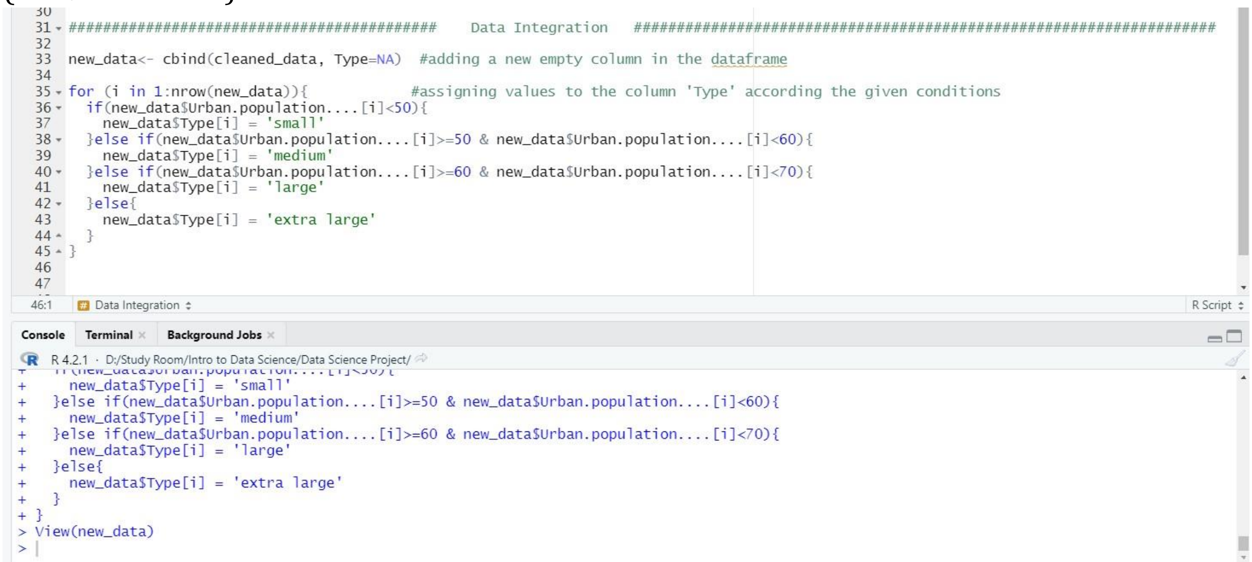
So, there were no more outliers in any column of the dataframe.

After Smoothing Noisy Data: (47 rows, 4 variables)

	States	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	182	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
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
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
33	North Carolina	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
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

From here we can see that, row 15, 32 and 40 were removed as there were noisy data.

2. **Data Integration:**
   In integration part, we have added a new column name "Type" to our latest cleaned dataframe, at first, we did not set any value for the column. Next, we used the given condition and assigned the values for each row of the Type column.
   The given condition was, small (<50%), medium (<60%), large (<70%), and extra-large (70% and above)



```r
############################################    Data Integration    ####################################################################

new_data<- cbind(cleaned_data, Type=NA)  #adding a new empty column in the dataframe

for (i in 1:nrow(new_data)){            #assigning values to the column 'Type' according the given conditions
  if(new_data$Urban.population....[i]<50){
    new_data$Type[i] = 'small'
  }else if(new_data$Urban.population....[i]>=50 & new_data$Urban.population....[i]<60){
    new_data$Type[i] = 'medium'
  }else if(new_data$Urban.population....[i]>=60 & new_data$Urban.population....[i]<70){
    new_data$Type[i] = 'large'
  }else{
    new_data$Type[i] = 'extra large'
  }
}
```

```
46:1   Data Integration                                                   R Script
Console   Terminal   Background Jobs
R 4.2.1 · D:/Study Room/Intro to Data Science/Data Science Project/
+     new_data$Type[i] = 'small'
+   }else if(new_data$Urban.population....[i]>=50 & new_data$Urban.population....[i]<60){
+     new_data$Type[i] = 'medium'
+   }else if(new_data$Urban.population....[i]>=60 & new_data$Urban.population....[i]<70){
+     new_data$Type[i] = 'large'
+   }else{
+     new_data$Type[i] = 'extra large'
+   }
+ }
> View(new_data)
>
```

   Here, we used a for loop and if-else condition to assign the values.

Before Data Integration: (47 rows, 4 variables)

	States	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	182	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
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
33	North Carolina	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
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

After Data Integration: (47 rows, 5 variables)

	States	Murder	Assault	Urban.population....	Type
1	Alabama	13.2	236	58	medium
2	Alaska	10.0	263	48	small
3	Arizona	8.1	294	80	extra large
4	Arkansas	8.8	190	50	medium
5	California	9.0	276	91	extra large
6	Colorado	7.9	204	78	extra large
7	Connecticut	3.3	110	77	extra large
8	Delaware	5.9	238	72	extra large
9	Florida	15.4	335	80	extra large
10	Georgia	17.4	182	60	large
11	Hawaii	5.3	46	83	extra large
12	Idaho	2.6	120	54	medium
13	Illinois	10.4	249	83	extra large
14	Indiana	7.2	113	65	large
16	Kansas	6.0	115	66	large
17	Kentucky	9.7	109	52	medium
18	Louisiana	15.4	249	66	large
19	Maine	2.1	83	51	medium
20	Maryland	11.3	300	67	large
21	Massachusetts	4.4	149	85	extra large
22	Michigan	12.1	255	74	extra large
23	Minnesota	2.7	72	66	large
24	Mississippi	16.1	259	44	small
25	Missouri	9.0	178	70	extra large

	States	Murder	Assault	Urban.population....	Type
24	Mississippi	16.1	259	44	small
25	Missouri	9.0	178	70	extra large
26	Montana	6.0	109	53	medium
27	Nebraska	4.3	102	62	large
28	Nevada	12.2	252	81	extra large
29	New Hampshire	2.1	57	56	medium
30	New Jersey	7.4	159	89	extra large
31	New Mexico	11.4	285	70	extra large
33	North Carolina	13.0	337	45	small
34	North Dakota	0.8	45	44	small
35	Ohio	7.3	120	75	extra large
36	Oklahoma	6.6	151	68	large
37	Oregon	4.9	159	67	large
38	Pennsylvania	6.3	106	72	extra large
39	Rhode Island	3.4	174	87	extra large
41	South Dakota	3.8	86	45	small
42	Tennessee	13.2	188	59	medium
43	Texas	12.7	201	80	extra large
44	Utah	3.2	120	80	extra large
45	Vermont	2.2	48	32	small
46	Virginia	8.5	156	63	large
47	Washington	4.0	145	73	extra large
48	West Virginia	5.7	81	39	small
49	Wisconsin	2.6	53	66	large
50	Wyoming	6.8	161	60	large

3. **Data Transformation:**
   Data transformation is a technique used to convert the raw data into a suitable format that efficiently eases data mining and retrieves strategic information. We have seen that the murders values are in decimal numbers, which is not logical. So, we transform the values of murder columns from decimal to integer which is known as numeric in R language.

```
########################################## Data Transformation ##############################

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


```

After Data Transformation: (47 rows, 5 variables)

	States	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	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
31	New Mexico	11	285	70	extra large
33	North Carolina	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
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

4. Data Discretization:

As we can see, all the attributes involved in our dataset are continuous type values in real numbers). However, depending on the model we want to build, we have to discretize the attribute values into binary or categorical types. For this dataset, we want to take the column Murders and Assaults values and categorize them based on the numbers of murder and assault. we will divide the risk factor of the areas in three categories, which are less crime, more crime, average crime. For example: murder number<=5 & assault number <=100 will define as 'less crime', murder number >5 and <=10 & assault number <=200 define as 'average crime' and the others will be 'more crime' area.

```r
########################################  Data Discretization ######################################################################

#based on the murder and assault cases we will divide the risk factor of the areas in three categories,
#which are less crime, more crime, average crime
#we will add a new column called danger_type

#murder number<=5 & assault number <=100 will define as less crime
#murder number >5 and <=10 & assault number <=200 define as average crime
#else the state will be more crime

discret_data<- cbind(new_data, danger_type=NA)  #adding a new column

for (i in 1:nrow(discret_data)){            #assigning values to the column 'danger_type' according to our classification
  if(discret_data$Murder[i]<=5 & discret_data$Assault[i]<=100)
  {
    discret_data$danger_type[i] = 'less crime'
  }else if(discret_data$Murder[i]>5 & discret_data$Murder[i]<=10 & discret_data$Assault[i]<=200){
    discret_data$danger_type[i] = 'average crime'
  }else{
    discret_data$danger_type[i] = 'more crime'
  }
}

final_data <- discret_data
final_data$Murder<- NULL
final_data$Assault<- NULL
```

Before Data Discretization:(47 rows, 5 variables)

	States	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	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
31	New Mexico	11	285	70	extra large
33	North Carolina	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
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

After Data Discretization:(47 rows, 4 variables)

	States	Urban.population....	Type	danger_type
1	Alabama	58	medium	more crime
2	Alaska	48	small	more crime
3	Arizona	80	extra large	more crime
4	Arkansas	50	medium	average crime
5	California	91	extra large	more crime
6	Colorado	78	extra large	more crime
7	Connecticut	77	extra large	more crime
8	Delaware	72	extra large	more crime
9	Florida	80	extra large	more crime
10	Georgia	60	large	more crime
11	Hawaii	83	extra large	less crime
12	Idaho	54	medium	more crime
13	Illinois	83	extra large	more crime
14	Indiana	65	large	average crime
16	Kansas	66	large	average crime
17	Kentucky	52	medium	average crime
18	Louisiana	66	large	more crime
19	Maine	51	medium	less crime
20	Maryland	67	large	more crime
21	Massachusetts	85	extra large	more crime
22	Michigan	74	extra large	more crime
23	Minnesota	66	large	less crime
24	Mississippi	44	small	more crime
25	Missouri	70	extra large	average crime
26	Montana	53	medium	average crime
27	Nebraska	62	large	more crime
28	Nevada	81	extra large	more crime
24	Mississippi	44	small	more crime
25	Missouri	70	extra large	average crime
26	Montana	53	medium	average crime
27	Nebraska	62	large	more crime
28	Nevada	81	extra large	more crime
29	New Hampshire	56	medium	less crime
30	New Jersey	89	extra large	average crime
31	New Mexico	70	extra large	more crime
33	North Carolina	45	small	more crime
34	North Dakota	44	small	less crime
35	Ohio	75	extra large	average crime
36	Oklahoma	68	large	average crime
37	Oregon	67	large	more crime
38	Pennsylvania	72	extra large	average crime
39	Rhode Island	87	extra large	more crime
41	South Dakota	45	small	less crime
42	Tennessee	59	medium	more crime
43	Texas	80	extra large	more crime
44	Utah	80	extra large	more crime
45	Vermont	32	small	less crime
46	Virginia	63	large	average crime
47	Washington	73	extra large	more crime
48	West Virginia	39	small	average crime
49	Wisconsin	66	large	less crime
50	Wyoming	60	large	average crime

Discussion and Conclusion:

We now have a clean dataset because all the procedures have been completed. Outliers and noisy data are no longer present. The mean value for the category has been used to fill in the gaps left by the missing data. In order to make it easier to interpret the data, we additionally do discretization and add a new category that deals with the data's range. When implementing the dataset, the data can be used.