1. 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.

	Murder	Assault	Urban Population (%)
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
Connecticut	3.3	110	77
Delaware	5.9	238	72
Florida	15.4	335	80
Georgia	17.4		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
Massachusetts	4.4	149	85
Michigan	12.1	255	74
Minnesota	2.7	72	66
Mississippi	16.1	259	44
Missouri	9	178	70
Montana	6	109	53
Nebraska	4.3	102	62
Nevada	12.2	252	81
New Hampshire	2.1	57	56
New Jersey	7.4	159	89
New Mexico	11.4	285	70
New York	11.1	254	6
North Carolina	13	337	45
North Dakota	0.8	45	44
Ohio	7.3	120	75

	Murder	Assault	Urban Population (%)
Oklahoma	6.6	151	68
Oregon	4.9	159	67
Pennsylvania	6.3	106	72
Rhode Island	3.4	174	87
South Carolina	14.4	879	48
South Dakota	3.8	86	45
Tennessee	13.2	188	59
Texas	12.7	201	80
Utah	3.2	120	80
Vermont	2.2	48	32
Virginia	8.5	156	63
Washington	4	145	73
West Virginia	5.7	81	39
Wisconsin	2.6	53	66
Wyoming	6.8	161	60

2. Project Solution Design:

The objective of this project is to preprocess the dataset in order to prepare it for data analysis. Initially, a data frame was created by combining four vectors, namely "State", "Murder", "Assault", and "UrbanPopulation". The first step of data preprocessing involved replacing the missing data with the mean value of the respective column. Further, to smoothen the noisy data, the outliers were replaced with more appropriate values. In the data reduction phase, double or floating-point numbers were converted to integers. Subsequently, in the first step of data discretization, a categorical column was created based on the "UrbanPopulation" variable. In the second step, an ordered factor variable named "OrderedFactorPopulation" was created based on the "PopulationLevel" variable. These steps have been followed to ensure the dataset is in an appropriate format for data analysis.

3. Data Frame:

```
State <- c ("Alabama","Alaska","Arizona","Arkansas","California","Colorado","Connecticut","Delaware",
            "Florida","Georgia","Hawaii","Idaho","Illinois","Indiana","Iowa","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","Wisconsin","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,NA,46,120,249,113,56,115,109,249,83,300,149,255,
              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)

UrbanPopulation <- 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)

US_State_Arrest_Stats_1973 <- data.frame(Murder,Assault,UrbanPopulation,row.names =State)

US_State_Arrest_Stats_1973
```

Figure: Code for Creating Data Frame

	Murder	Assault	UrbanPopulation
Alabama	13.2	236	58
Alaska	10.0	263	48
Arizona	8.1	294	80
Arkansas	8.8	190	50
California	9.0	276	91
Colorado	7.9	204	78
Connecticut	3.3	110	77
Delaware	5.9	238	72
Florida	15.4	335	80
Georgia	17.4	*NA*	60
Hawaii	5.3	46	83
Idaho	2.6	120	54
Illinois	10.4	249	83
Indiana	7.2	113	65
Iowa	2.2	56	570
Kansas	6.0	115	66

Kentucky	9.7	109	52
Louisiana	15.4	249	66
Maine	2.1	83	51
Maryland	11.3	300	67
Massachusetts	4.4	149	85
Michigan	12.1	255	74
Minnesota	2.7	72	66
Mississippi	16.1	259	44
Missouri	9.0	178	70
Montana	6.0	109	53
Nebraska	4.3	102	62
Nevada	12.2	252	81
New Hampshire	2.1	57	56
New Jersey	7.4	159	89
New Mexico	11.4	285	70
New York	11.1	254	6
North Carolina	13.0	337	45
North Dakota	0.8	45	44
Ohio	7.3	120	75
Oklahoma	6.6	151	68
Oregon	4.9	159	67
Pennsylvania	6.3	106	72
Rhode Island	3.4	174	87
South Carolina	14.4	879	48
South Dakota	3.8	86	45
Tennessee	13.2	188	59
Texas	12.7	201	80
Utah	3.2	120	80
Vermont	2.2	48	32
Virginia	8.5	156	63
Washington	4.0	145	73

West Virginia	5.7	81	39
Wisconsin	2.6	53	66
Wyoming	6.8	161	60

Figure: Initial Data Frame (US_State_Arrest_Stats_1973)

4. Data Preprocessing:

4.1. Data cleaning:

Handling Missing Data:

Georgia	17.4	*NA*	60

We have a missing data at the “Assault” column of “Georgia” state. In order to replace the missing data, first it was replaced with 0 then we took the mean value of “Assault” column and replaced the value.

```
US_State_Arrest_Stats_1973$Assault[is.na(US_State_Arrest_Stats_1973$Assault)] <- 0

assaultMean <- mean(US_State_Arrest_Stats_1973$Assault)

US_State_Arrest_Stats_1973$Assault[US_State_Arrest_Stats_1973$Assault == 0] <- assaultMean

US_State_Arrest_Stats_1973
```

Figure: Code for Replacing the Missing Data

Georgia	17.4	178.54	60

Figure: After Replacing the Missing Data

Smooth Noisy Data:

In order to find the noisy data or the outlier we used order function to sort the data.

```
Murder_sorted <- US_State_Arrest_Stats_1973[order(US_State_Arrest_Stats_1973$Murder), "Murder"]
Murder_sorted

Assault_sorted <- US_State_Arrest_Stats_1973[order(US_State_Arrest_Stats_1973$Assault), "Assault"]
Assault_sorted

UrbanPopulation_sorted <- US_State_Arrest_Stats_1973[order(US_State_Arrest_Stats_1973$UrbanPopulation), "UrbanPopulation"]
UrbanPopulation_sorted
```

Figure: Code for Order Function to Find Outliers

```
> Murder_sorted <- US_State_Arrest_Stats_1973[order(US_State_Arrest_Stats_1973$Murder), "Murder"]
> Murder_sorted
 [1]  0.8  2.1  2.1  2.2  2.2  2.6  2.6  2.7  3.2  3.3  3.4  3.8  4.0  4.3  4.4  4.9  5.3  5.7  5.9  6.0  6.0
[22]  6.3  6.6  6.8  7.2  7.3  7.4  7.9  8.1  8.5  8.8  9.0  9.0  9.7 10.0 10.4 11.1 11.3 11.4 12.1 12.2 12.7
[43] 13.0 13.2 13.2 14.4 15.4 15.4 16.1 17.4
>
> Assault_sorted <- US_State_Arrest_Stats_1973[order(US_State_Arrest_Stats_1973$Assault), "Assault"]
> Assault_sorted
 [1]  45.00  46.00  48.00  53.00  56.00  57.00  72.00  81.00  83.00  86.00 102.00 106.00 109.00 109.00 110.00
[16] 113.00 115.00 120.00 120.00 120.00 145.00 149.00 151.00 156.00 159.00 159.00 161.00 174.00 178.00 178.54
[31] 188.00 190.00 201.00 204.00 236.00 238.00 249.00 249.00 252.00 254.00 255.00 259.00 263.00 276.00 285.00
[46] 294.00 300.00 335.00 337.00 879.00
>
> UrbanPopulation_sorted <- US_State_Arrest_Stats_1973[order(US_State_Arrest_Stats_1973$UrbanPopulation), "UrbanPopulation"]
> UrbanPopulation_sorted
 [1]   6  32  39  44  44  45  45  48  48  50  51  52  53  54  56  58  59  60  60  62  63  65  66  66  66  66  67
[28]  67  68  70  70  72  72  73  74  75  77  78  80  80  80  80  81  83  83  85  87  89  91 570
>
```

Figure: Output of Order Function to Find Outliers

In the “Murder” column we don’t see any abnormal value which can be identified as outlier but in the “Assault” column we see value 879.00 which is pretty abnormal compared to other values of that column, same for the “Urban Population” column the value 6 is significantly low and 570 is significantly higher compared to other values.

South Carolina	14.4	879.00	48
New York	11.1	254.00	6
Iowa	2.2	56.00	570

Figure: Sample with Outliers

To address the problem of outlier we have replace the value of “Assault” column 879.00 using the mean of the “Assault” column values. For the case of “UrbanPopulation” column the value 6 and 570 we assume there is some missing input for the value 6, so we replace 6 with 60 and for 570 we assume the 0 is extra since it cannot be more than 100 so we replace 570 with 57.

```
US_State_Arrest_Stats_1973$Assault[US_State_Arrest_Stats_1973$Assault == 879] <- assaultMean

US_State_Arrest_Stats_1973$UrbanPopulation[US_State_Arrest_Stats_1973$UrbanPopulation == 6] <- 60

US_State_Arrest_Stats_1973$UrbanPopulation[US_State_Arrest_Stats_1973$UrbanPopulation == 570] <- 57
```

Figure: Code for Replacing Outliers

South Carolina	14.4	178.54	48
New York	11.1	254.00	60
Iowa	2.2	56.00	57

Figure: After Replacing Outliers

Data Wrangling or Munging:

Fortunately for this dataset no wrangling or munging is needed since all data are in same pattern.

4.2. Data Integration:

For this project no data integration is needed as we collected the data from a single source.

4.3. Data Transformation:

The steps of data transformation are already performed during smoothing noisy data. So, transformation is not required anymore.

4.4. Data Reduction:

The values of “Assault” column was integer value before replacing the missing value with the mean value but now they are float or decimal so we can reduce the size by ceiling those value with no value after the decimal point.

	Murder	Assault	UrbanPopulation
Alabama	13.2	236.00	58
Alaska	10.0	263.00	48
Arizona	8.1	294.00	80
Arkansas	8.8	190.00	50
California	9.0	276.00	91
Colorado	7.9	204.00	78
Connecticut	3.3	110.00	77
Delaware	5.9	238.00	72
Florida	15.4	335.00	80
Georgia	17.4	178.54	60

```
US_State_Arrest_Stats_1973$Assault <- ceiling(US_State_Arrest_Stats_1973$Assault)
```

	Murder	Assault	UrbanPopulation
Alabama	13.2	236	58
Alaska	10.0	263	48
Arizona	8.1	294	80
Arkansas	8.8	190	50
California	9.0	276	91
Colorado	7.9	204	78
Connecticut	3.3	110	77
Delaware	5.9	238	72
Florida	15.4	335	80
Georgia	17.4	179	60

Figure: Before and After the Data Reduction of Assault Column Using Ceiling Function

Similarly, for the "Murder" column we used ceiling function if the value is greater or equal to 5 after the decimal point and used floor function if the value is less than 5 after the decimal point.

	Murder	Assault	UrbanPopulation
Alabama	13.2	236	58
Alaska	10.0	263	48
Arizona	8.1	294	80
Arkansas	8.8	190	50
California	9.0	276	91
Colorado	7.9	204	78
Connecticut	3.3	110	77
Delaware	5.9	238	72
Florida	15.4	335	80
Georgia	17.4	179	60

```
US_State_Arrest_Stats_1973$Murder <- ifelse((US_State_Arrest_Stats_1973$Murder %% 1) >= 0.5,
                                            ceiling(US_State_Arrest_Stats_1973$Murder), floor(US_State_Arrest_Stats_1973$Murder))
```

	Murder	Assault	UrbanPopulation
Alabama	13	236	58
Alaska	10	263	48
Arizona	8	294	80
Arkansas	9	190	50
California	9	276	91
Colorado	8	204	78
Connecticut	3	110	77
Delaware	6	238	72
Florida	15	335	80
Georgia	17	179	60

Figure: Before and After the Data Reduction of Murder Column Using Ceiling and Floor Function

4.5. Data Discretization:

In this part of the project, we added two new columns. We created the first column "PopulationLevel" with the help of "UrbanPopulation". Convert the urban population percentage into level, small (<50%), medium (>= 50% to <60%), large (>= 60 to <70%), and extra-large (70% and above) to create the column with categorical data.

```
PopulationLevel <- ifelse(US_State_Arrest_Stats_1973$UrbanPopulation < 50, "Small",
                   ifelse(US_State_Arrest_Stats_1973$UrbanPopulation < 60, "Medium",
                   ifelse(US_State_Arrest_Stats_1973$UrbanPopulation < 70, "Large", "Extra-Large")))

US_State_Arrest_Stats_1973$PopulationLevel <- PopulationLevel
```

	Murder	Assault	UrbanPopulation	PopulationLevel
Alabama	13	236	58	Medium
Alaska	10	263	48	Small
Arizona	8	294	80	Extra-Large
Arkansas	9	190	50	Medium
California	9	276	91	Extra-Large
Colorado	8	204	78	Extra-Large
Connecticut	3	110	77	Extra-Large
Delaware	6	238	72	Extra-Large
Florida	15	335	80	Extra-Large
Georgia	17	179	60	Large

Figure: First 10 Rows After Adding PopulationLevel Column

	State	Murder	Assault	UrbanPopulation	PopulationLevel
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	179	60	Large

For the second column. Since, the data of the column “PopulationLevel” are categorical, we create another column named “OrderedFactorPopulation” based on “PopulationLevel” which is an ordered factor variable. Where, Small = 1, Medium = 2, Large = 3, and Extra-large = 4.

```
OrderedFactorPopulation <- factor(PopulationLevel,levels <-c("Small","Medium","Large","Extra-Large"),labels <-c(1,2,3,4))

US_State_Arrest_Stats_1973$OrderedFactorPopulation <- OrderedFactorPopulation
```

	Murder	Assault	UrbanPopulation	PopulationLevel	OrderedFactorPopulation
Alabama	13	236	58	Medium	2
Alaska	10	263	48	Small	1
Arizona	8	294	80	Extra-Large	4
Arkansas	9	190	50	Medium	2
California	9	276	91	Extra-Large	4
Colorado	8	204	78	Extra-Large	4
Connecticut	3	110	77	Extra-Large	4
Delaware	6	238	72	Extra-Large	4
Florida	15	335	80	Extra-Large	4
Georgia	17	179	60	Large	3

Figure: First 10 Rows After Adding OrderedFactorPopulation Column

We change some of the column's name to match the given name in the project with colnames function.

```
colnames(US_State_Arrest_Stats_1973)[4] <- "Urban Population (%)"
colnames(US_State_Arrest_Stats_1973)[5] <- "Population Level"
colnames(US_State_Arrest_Stats_1973)[6] <- "Ordered Factor Population"
```

The Cleaned Dataset:

	Murder	Assault	Urban Population (%)	Population Level	Ordered Factor Population
Alabama	13	236	58	Medium	2
Alaska	10	263	48	Small	1
Arizona	8	294	80	Extra-Large	4
Arkansas	9	190	50	Medium	2
California	9	276	91	Extra-Large	4
Colorado	8	204	78	Extra-Large	4
Connecticut	3	110	77	Extra-Large	4
Delaware	6	238	72	Extra-Large	4
Florida	15	335	80	Extra-Large	4
Georgia	17	179	60	Large	3
Hawaii	5	46	83	Extra-Large	4
Idaho	3	120	54	Medium	2
Illinois	10	249	83	Extra-Large	4
Indiana	7	113	65	Large	3
Iowa	2	56	57	Medium	2
Kansas	6	115	66	Large	3
Kentucky	10	109	52	Medium	2
Louisiana	15	249	66	Large	3
Maine	2	83	51	Medium	2
Maryland	11	300	67	Large	3
Massachusetts	4	149	85	Extra-Large	4
Michigan	12	255	74	Extra-Large	4
Minnesota	3	72	66	Large	3
Mississippi	16	259	44	Small	1
Missouri	9	178	70	Extra-Large	4

Montana	6	109	53	Medium	2
Nebraska	4	102	62	Large	3
Nevada	12	252	81	Extra-Large	4
New Hampshire	2	57	56	Medium	2
New Jersey	7	159	89	Extra-Large	4
New Mexico	11	285	70	Extra-Large	4
New York	11	254	60	Large	3
North Carolina	13	337	45	Small	1
North Dakota	1	45	44	Small	1
Ohio	7	120	75	Extra-Large	4
Oklahoma	7	151	68	Large	3
Oregon	5	159	67	Large	3
Pennsylvania	6	106	72	Extra-Large	4
Rhode Island	3	174	87	Extra-Large	4
South Carolina	14	179	48	Small	1
South Dakota	4	86	45	Small	1
Tennessee	13	188	59	Medium	2
Texas	13	201	80	Extra-Large	4
Utah	3	120	80	Extra-Large	4
Vermont	2	48	32	Small	1
Virginia	9	156	63	Large	3
Washington	4	145	73	Extra-Large	4
West Virginia	6	81	39	Small	1
Wisconsin	3	53	66	Large	3
Wyoming	7	161	60	Large	3

Figure: Final Dataset After Preprocessing