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.

	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
Iowa	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

	Murder	Assault	Urban population (%)
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
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

At first, **assign** this dataset into a data frame in R.

Now, you have to **apply the pre-processing techniques** to prepare the dataset for data analysis. To prepare a cleaned dataset, perform the following tasks of data pre-processing using R language:

1. Data cleaning:
    a. Smooth Noisy Data
    b. Handling Missing Data
    c. Data Wrangling or Munging
2. Data Integration
3. Data Transformation
4. Data Reduction
5. Data Discretization

Next, add a new column (named **population level**) in the data frame based on the **urban population** variable. [Hint: Convert the urban population percentage into level, for example, small (<50%), medium (>= 50% to <60%), large (>= 60 to <70%), and extra-large (70% and above).]

As the newly added variable (i.e., **population level**) are numerical, so **convert** the **population level** variable into an ordered factor variable (named **ordered factor population**) like small = 1, medium = 2, large = 3, and extra-large = 4, add it to the data frame. So, in total the data frame will contain 6 variables.

Finally, **display** the full cleaned dataset.

Project Report:

You have to submit a complete report mentioning every step of data pre-processing in detail.

The project report should contain the following information:

- A standard cover-page
- 1. Project overview: Write details about the project
- 2. Project solution design: Write how do you design the solution to complete this project
- 3. Data Frame: Create the data frame from the dataset.
- 3. Data pre-processing: Detailed description of each tasks of data pre-processing to prepare a cleaned dataset along with R code for the project with output (attach screenshot of the output where necessary with description).
- New variable integration: Describe how do you add two new variables (as discussed in the project description).
- The Cleaned Dataset: Display the full cleaned dataset.

Remember that you have to submit printed version of your project report and face the viva to complete the project.