1. Project overview:

In the real world, data is frequently dirty; here dirty data means incomplete, noisy or inconsistent data. Before using data for a specific purpose, it must first be cleaned up. As data generation is expanding quickly these days and there are more and more heterogeneous data sources, the chance of collecting anomalous or erroneous data is relatively considerable. That's why data pre-processing is essential before its actual use. Having clean data will ultimately increase overall productivity and allow for the highest quality information in decision-making. Data pre-processing is the concept of changing the raw data into a clean data set. The dataset is pre-processed in order to check missing values, noisy data, and other inconsistencies before executing it to the algorithm.

To perform data analysis on the following dataset, this is needed to pre-process data because the project dataset may contain noisy data, missing values and errors or outliers. The following dataset is based on the statistics in arrests per 100,000 residents for assault and murder, in each of the 50 US states, in 1973. The percentage of the population living in urban areas is also given.

To prepare a cleaned dataset, it is needed to perform the following tasks of data pre-processing using R language:

- **i.** Data cleaning:
  - **a.** Smooth Noisy Data
  - **b.** Handling Missing Data
  - **c.** Data Wrangling or Munging
- **ii.** Data Integration
- **iii.** Data Transformation
- **iv.** Data Reduction
- **v.** Data Discretization

2. Project solution design:

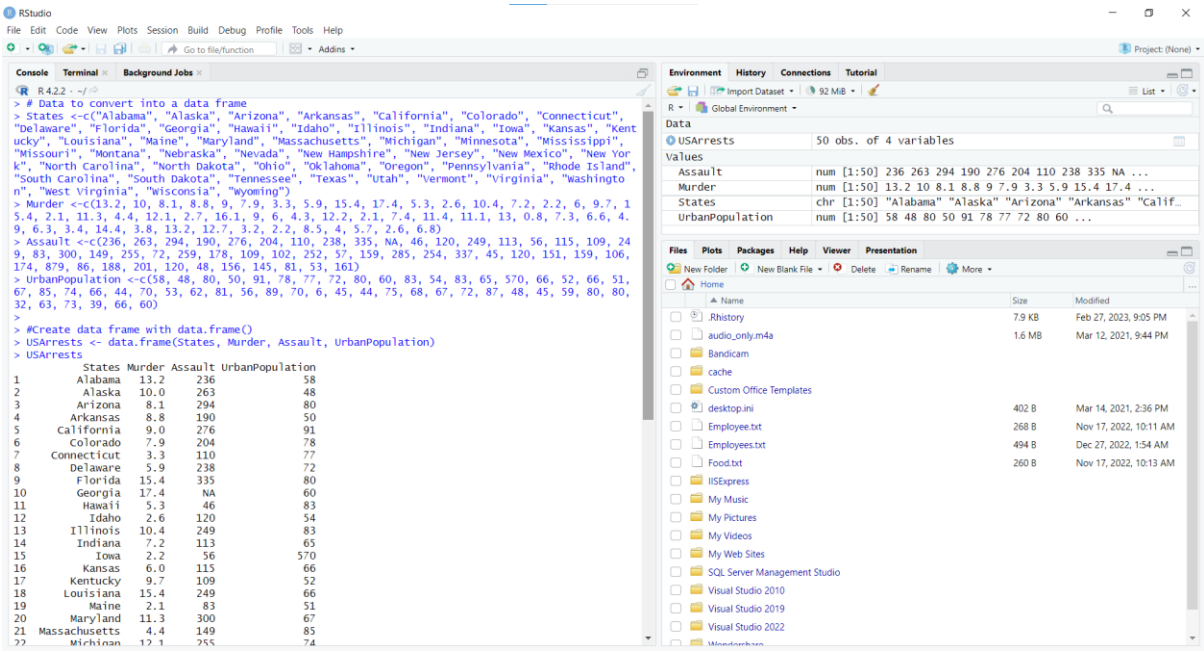
- Load all the data from the pdf file by using data.frame() function.
- Check the dataset for the missing values. If there is a missing value, then that will be replaced with the mean value.
- Check the dataset for smooth noisy data. If there is a noisy data, then replace it with the mean value.
- Add a new column named Type to classify the population according to their size.
- And last save the cleaned data.

Tools Used:

I. RStudio
II. Word

3. Data Frame:

Dataset into a data frame of the project:



```
> # Data to convert into a data frame
> States <-c("Alabama", "Alaska", "Arizona", "Arkansas", "California", "Colorado", "Connecticut",
"Delaware", "Florida", "Georgia", "Hawaii", "Idaho", "Illinois", "Indiana", "Iowa", "Kansas", "Kent
ucky", "Louisiana", "Maine", "Maryland", "Massachusetts", "Michigan", "Minnesota", "Mississippi",
"Missouri", "Montana", "Nebraska", "Nevada", "New Hampshire", "New Jersey", "New Mexico", "New Yor
k", "North Carolina", "North Dakota", "Ohio", "Oklahoma", "Oregon", "Pennsylvania", "Rhode Island",
"South Carolina", "South Dakota", "Tennessee", "Texas", "Utah", "Vermont", "Virginia", "Washingto
n", "West Virginia", "Wisconsia", "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, 1
5.4, 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, 24
9, 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)
>
> #Create data frame with data.frame()
> USArrests <- data.frame(States, Murder, Assault, UrbanPopulation)
> USArrests
          States Murder Assault UrbanPopulation
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
```

```
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     Wisconsia    2.6      53             66
50       Wyoming    6.8     161             60
```

4. Data pre-processing:

1) Data cleaning:

Data cleaning is the process of removing incorrect, duplicate or otherwise erroneous data from a dataset. These errors can include incorrectly formatted data, redundant entries, mislabeled data, and other issues.

Handling Missing Data: If there have missing data in dataset, there are several ways to handle it in R programming. One way is to simply remove any rows or columns that contain missing data. Another way to handle missing data is to impute the missing values using a statistical method.

This dataset contains missing values in the assault variable. In R programming the missing value will be undefiled and with undefiled, any arithmetic operation will produce a NAN. So, we have to replace these missing values with the mean values of the respective variables.

By using the "sum(is.na(column_name))" method, missing data can be checked.

```
> # Data Preprocessing
> # Handling Missing Data:
> sum(is.na(USArrests$States))
[1] 0
> sum(is.na(USArrests$Murder))
[1] 0
> sum(is.na(USArrests$Assault))
[1] 1
> sum(is.na(USArrests$UrbanPopulation))
[1] 0
```

Replace missing values with the mean values of the respective variables:

```
> USArrests$Assault[is.na(USArrests$Assault)]<- mean(USArrests$Assault,na.rm = TRUE)
> print(USArrests)
           States Murder  Assault UrbanPopulation
1         Alabama   13.2 236.0000              58
2          Alaska   10.0 263.0000              48
3         Arizona    8.1 294.0000              80
4        Arkansas    8.8 190.0000              50
5      California    9.0 276.0000              91
6        Colorado    7.9 204.0000              78
7     Connecticut    3.3 110.0000              77
8        Delaware    5.9 238.0000              72
9         Florida   15.4 335.0000              80
10        Georgia   17.4 182.1837              60
11         Hawaii    5.3  46.0000              83
12          Idaho    2.6 120.0000              54
13       Illinois   10.4 249.0000              83
14        Indiana    7.2 113.0000              65
15           Iowa    2.2  56.0000             570
16         Kansas    6.0 115.0000              66
17       Kentucky    9.7 109.0000              52
18      Louisiana   15.4 249.0000              66
19          Maine    2.1  83.0000              51
20       Maryland   11.3 300.0000              67
21  Massachusetts    4.4 149.0000              85
22       Michigan   12.1 255.0000              74
23      Minnesota    2.7  72.0000              66
24    Mississippi   16.1 259.0000              44
25       Missouri    9.0 178.0000              70
26        Montana    6.0 109.0000              53
27       Nebraska    4.3 102.0000              62
28         Nevada   12.2 252.0000              81
29  New Hampshire    2.1  57.0000              56
30     New Jersey    7.4 159.0000              89
31     New Mexico   11.4 285.0000              70
32       New York   11.1 254.0000               6
33 North Carolina   13.0 337.0000              45
34   North Dakota    0.8  45.0000              44
35           Ohio    7.3 120.0000              75
36       Oklahoma    6.6 151.0000              68
37         Oregon    4.9 159.0000              67
38   Pennsylvania    6.3 106.0000              72
```

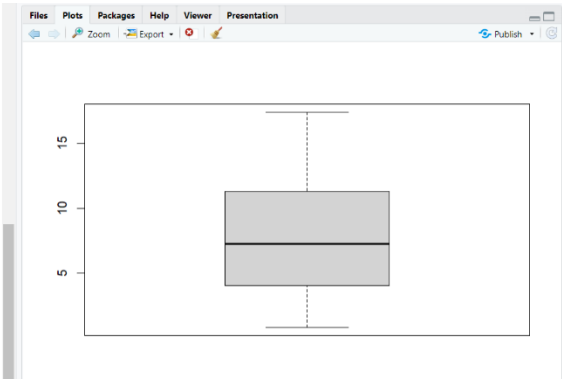
```
39   Rhode Island    3.4 174.0000              87
40 South Carolina   14.4 879.0000              48
41   South Dakota    3.8  86.0000              45
42      Tennessee   13.2 188.0000              59
43          Texas   12.7 201.0000              80
44           Utah    3.2 120.0000              80
45        Vermont    2.2  48.0000              32
46       Virginia    8.5 156.0000              63
47     Washington    4.0 145.0000              73
48  West Virginia    5.7  81.0000              39
49      Wisconsia    2.6  53.0000              66
50        Wyoming    6.8 161.0000              60
>
```

Smooth Noisy Data:

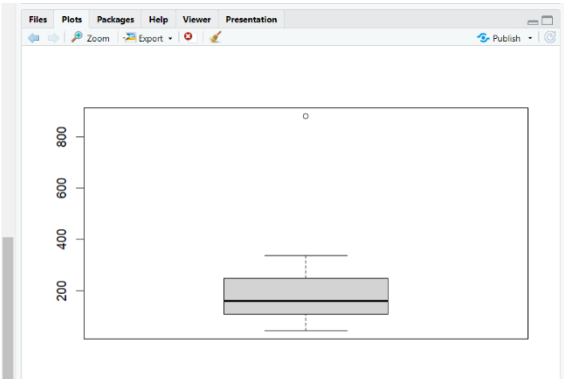
Noisy data are data with a large amount of additional meaningless information in it called noise. noise can be defined as mislabeled examples (class noise) or errors or outliers in the values of attributes (attribute noise)

By using the boxplot() method, we can detect the outliers.

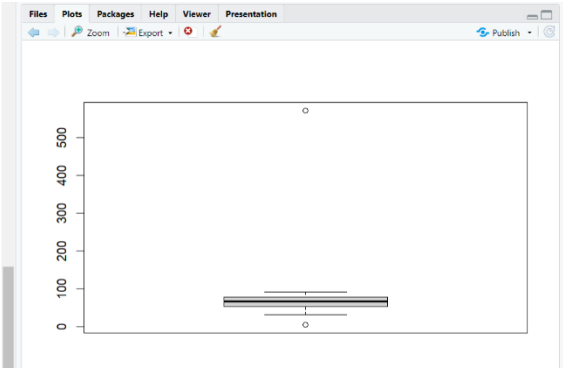
```
> #Smooth Noisy Data:
> boxplot(USArrests$Murder)
```




```
> boxplot(USArrests$Assault)
```



```
> boxplot(USArrests$UrbanPopulation)
```



The Assault and Urban Population columns contain outliers.

```
> # Data is replacing by mean value
> USArrests$Assault[USArrests$Assault == 879.0000] <- mean(USArrests$Assault)
> USArrests$UrbanPopulation[USArrests$UrbanPopulation == 570] <- 57
> USArrests$UrbanPopulation[USArrests$UrbanPopulation == 6] <- 60
```

2) Data Integration:

Data integration is the process of combining data from different sources to help data managers and executives analyze it and make smarter business decisions.

There is no need to do data integration. Because there is no other dataset on this project.

3) Data Transformation:

Data transformation is the process of converting, cleansing, and structuring data into a usable format that can be analyzed to support decision making processes, and to propel the growth of an organization. Data transformation is used when data needs to be converted to match that of the destination system.

Converting the Murder and Assault values into integers:

```
> # Data Transformation
> #values of murder and assault are in decimal which is not possible; converting them into integers
> USArrests$Murder = as.numeric(format(round(USArrests$Murder, 0)))
> USArrests$Assault = as.numeric(format(round(USArrests$Assault, 0)))
> print(USArrests)
           States Murder Assault UrbanPopulation
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       Illinois     10     249              83
14        Indiana      7     113              65
15           Iowa      2      56              57
16         Kansas      6     115              66
17       Kentucky     10     109              52
18      Louisiana     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 Hampshire      2      57              56
30     New Jersey      7     159              89
31     New Mexico     11     285              70
32       New York     11     254              60
33 North Carolina     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 Carolina     14     182              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      Wisconsia      3      53              66
50        Wyoming      7     161              60
>
```

4) Data Reduction:

Data reduction is a capacity optimization technique in which data is reduced to its simplest possible form to free up capacity on a storage device. Data reduction reduces the amount of data that is stored on the system using a number of methods. When dealing with high dimensional data, it is often useful to reduce the dimensionality by projecting the data to a lower dimensional subspace which captures the “essence” of the data.

As the following dataset is not so large therefore, there is no need to do data reduction.

5) **Data Discretization:**

Data discretization is defined as a process of converting continuous data attribute values into a finite set of intervals and associating with each interval some specific data value. The goal of discretization is to reduce the number of values a continuous variable assumes by grouping them.

For the following dataset, data discretization is not needed.

5. New variable integration:

Preparing the dataset to integrate a new column (named type) based on the Urban Population variable. Convert the urban population percentage into types. For example: small (<50%), medium (<60%), large (<70%), and extra-large (70% and above).

```
> library(dplyr)
> PopululationLevel <- USArrests %>% mutate(
+     Type = case_when(
+         UrbanPopulation <50 ~ "Small",
+         UrbanPopulation < 60 ~ "Medium",
+         UrbanPopulation < 70 ~ "Large",
+         UrbanPopulation >= 70 ~ "Extra Large"
+     )
+ )
> print(PopululationLevel)
           States Murder Assault UrbanPopulation        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
15           Iowa      2      56              57      Medium
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
```

```
32        New York     11     254              60        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
40 South Carolina     14     182              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      Wisconsia      3      53              66        Large
50        Wyoming      7     161              60        Large
>
```

Convert the population level variable into an ordered factor variable:

```
> OrderedFactorPopulation<-factor(PopululationLevel ,levels=c("Small","Medium","Large","Extra Large"),
+                                 labels=c(1,2,3,4))
> print(OrderedFactorPopulation)
```

6. The Cleaned Dataset:

By pre-processing the dataset using the R language, a cleaned dataset is generated. Now the data is ready for the analysis phase.