



AMERICAN INTERNATIONAL UNIVERSITY- BANGLADESH

Project Title: Apply Data Pre-processing on a Dataset

Submitted By:

Name: Sayem, Easinul Abedin

ID:19-40291-1

Section:[C]

Subject: INTRODUCTION TO DATA SCIENCE

Submitted To:

Teachers Name: DR.AKINUL ISLAM JONY

Submission Date:23/10/2022

Project overview:

This project is based on data pre-processing. A database may have missing data for a number of reasons, such as that the column needs a value or that the data was not captured at the time of data collection. The process of cleaning up raw data and turning it into usable information is known as data preparation. In this project, we will analyse the dataset contains statistics in arrests per 100,000 residents for assault and murder, in each of the 50 US states, in 1973. There are Also given is the percentage of the population living in urban areas.

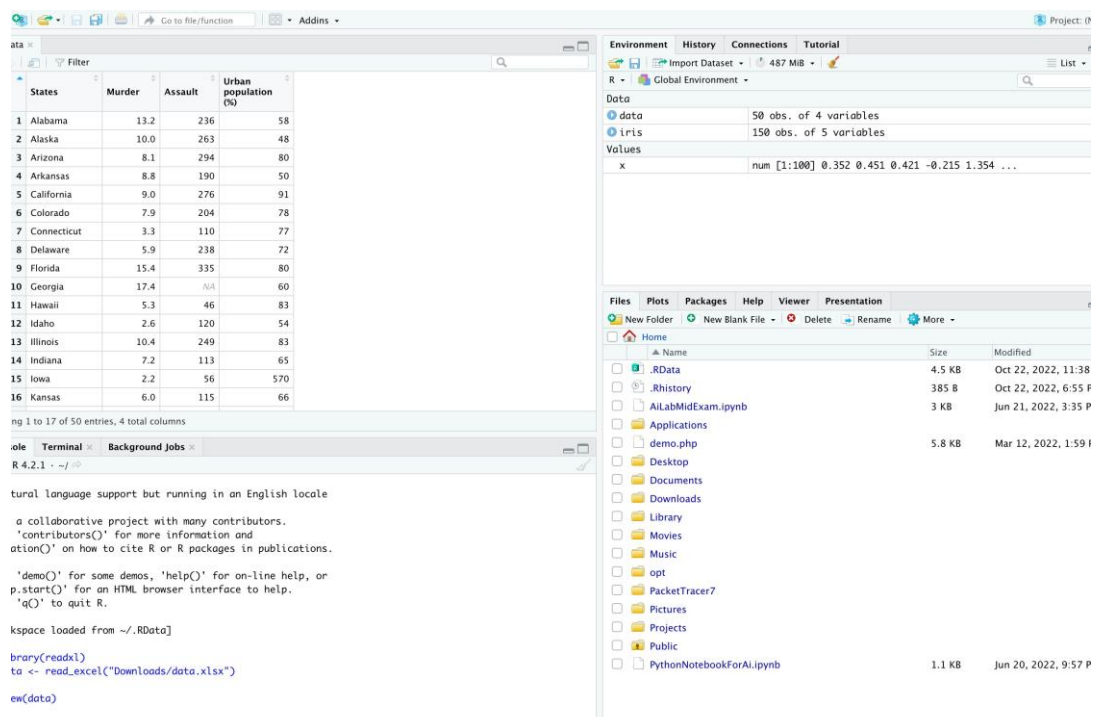
project solution design:

For read the excel file :

```
>library(readxl)
>data <- read_excel("Downloads /data.xlsx")
```

For View:

```
>View(data)
```



Read and View

For Fixed a data name:

```
>data[data=="iu7vh6"] <- "SAYEM"
```

The screenshot displays the RStudio environment. The main editor window shows a data frame with 16 rows and 4 columns: States, Murder, Assault, and Urban population (%). The data is as follows:

| 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 |

The Environment pane on the right shows the Global Environment with two objects: 'data' (50 obs. of 4 variables) and 'iris' (150 obs. of 5 variables). The Files pane at the bottom shows the file explorer with various files and folders.

```
R 4.2.1 ~ /  
tural language support but running in an English locale  
  
a collaborative project with many contributors.  
'contributors()' for more information and  
ation()' on how to cite R or R packages in publications.  
  
'demo()' for some demos, 'help()' for on-line help, or  
p.start()' for an HTML browser interface to help.  
'q()' to quit R.  
  
kspace loaded from ~/RData]  
  
brary(readxl)  
ta <- read_excel("Downloads/data.xlsx")  
  
ew(data)  
ta[data=="iu7vh6"] <- "Prajukta"
```

Fixed a Data name

Then step by step , Data Cleaning >Data Integration > Data Transformation >Data Reduction
>Data Discretization will be solved

Data pre-processing:

The process of cleaning up raw data and turning it into usable information is known as data preparation. The data can then be used for a variety of tasks, including reporting and decision-making.

Data pre-processing techniques are used when the data is inconsistent, which indicates that the data is not recorded in accordance with the constraints on the column, noisy, which may contain numerous errors or outliers, and incomplete, which indicates that some attribute value is missing.

The database may have missing data for a number of reasons, such as that the column needs a value to be entered or that the data was not captured at the time of data collection.

Several methods can be used to fill in the missing values:

The first approach to handling the missing value of Assault in place G ignores the tuple; the second approach fills the Assault value of place G with "unknown."

The data considered to be outliers are those that may differ from all other data or those that may result in problems. The percentage of the urban population in this case must range from 0% to 100%. The value of the urban population being greater than 100% is therefore unusual.

State I, which has 570 people living in metropolitan areas, is an anomaly in the data set, as is state G, which lists assault as "unknown."

Take away the rows to smooth the data and get rid of these two outliers.

The urban population is split into four categories: small for those with less than 50% of the total, medium for those with between 50% and 60%, big for those with between 60% and 70%, and extra-large for those with more than 70% of the total.

Most serious, less serious, and least serious crimes are categorized similarly. The most serious crimes go into three categories: murder over 10, assault over 250, and assault between 5 and 10, between 110 and 250. The least serious crimes fall into the categories of murder under 5, and assault under 110.

Data Cleaning:

Data Munging:

No munging or wrangling is necessary for the data set.

Handling Missing Data:

```
>mean(data$Assault, na.rm=TRUE)
```

The screenshot displays the RStudio environment. The main editor shows a data frame with 50 rows and 4 columns: States, Murder, Assault, and Urban population (%). The data is sorted by the 'Assault' variable. The environment pane on the right shows the 'data' object with 50 observations of 4 variables. The file explorer on the bottom right shows the project structure, including a 'data' folder and various files like '.RData', '.Rhistory', and 'demo.php'.

| States | Murder | Assault | Urban population (%) |
|----------------|--------|---------|----------------------|
| Oklahoma | 6.6 | 15.1 | 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 |

```
mean(data$Assault, na.rm=TRUE)
[1] 182.18
```

Missing Handling

```
>data[is.na(data)] <- mean(data$Assault, na.rm=TRUE)
```

The screenshot displays the RStudio environment. The main editor window shows a data frame with the following structure:

| | States | Murder | Assault | Urban population (%) |
|----|-------------|--------|---------|----------------------|
| 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 | 570 |
| 16 | Kansas | 6 | 115 | 66 |

The R console at the bottom shows the command:

```
data[is.na(data)] <- mean(data$Assault, na.rm=TRUE)
```

The Environment pane on the right shows the following objects:

- `data`: 50 obs. of 4 variables
- `iris`: 150 obs. of 5 variables

The Files pane on the right shows the following files:

- `.RData`: 4.5 KB, Oct 22, 2022, 11:38
- `.Rhistory`: 385 B, Oct 22, 2022, 6:55 F
- `AI.LabMidExam.ipynb`: 3 KB, Jun 21, 2022, 3:35 P
- `Applications`
- `demo.php`: 5.8 KB, Mar 12, 2022, 1:59 I
- `Desktop`
- `Documents`
- `Downloads`
- `Library`
- `Movies`
- `Music`
- `opt`
- `PacketTracer7`
- `Pictures`
- `Projects`
- `Public`
- `PythonNotebookForAI.ipynb`: 1.1 KB, Jun 20, 2022, 9:57 P

Missing handling

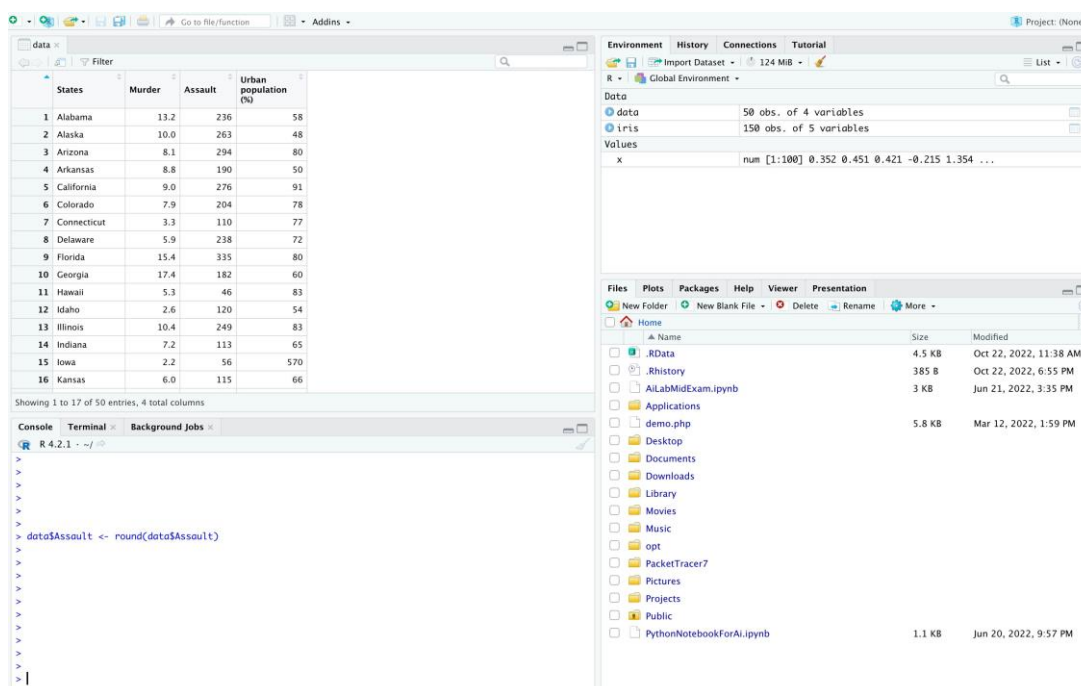
Smooth Noisy Data:

Since the number of people who have died cannot be decimal, the data in the Murder column is noisy.

Iowa 570, South Carolina, is a unique case in the urban population (%) of 879.

For assault :

```
>data$Assault <- round(data$Assault)
```



The screenshot displays the RStudio environment. The main window shows a data frame with the following data:

| 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 |

The console shows the command: `> data$Assault <- round(data$Assault)`

Smooth Noisy Data for Assault

For Murder :

```
>data$Murder <- round (data$Murder)
```

The screenshot displays the RStudio environment. The main editor window shows a data frame with the following columns: States, Murder, Assault, and Urban population (%). The data is sorted by the Murder rate in descending order. The Environment pane on the right shows the 'data' object with 50 observations of 4 variables. The Files pane on the bottom right shows the project directory structure, including files like .RData, .Rhistory, and various subdirectories like Applications, Desktop, Documents, Downloads, Library, Movies, Music, opt, PacketTracer7, Pictures, Projects, Public, and PythonNotebookForAi.ipynb.

| States | Murder | Assault | Urban population (%) |
|---------------|--------|---------|----------------------|
| 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 | 570 |
| 16 Kansas | 6 | 115 | 66 |

```
data$Murder <- round (data$Murder)
```

Smooth Noisy Data for Murder

Data Integration:

We estimate the overall number of rape murders in India as the average number of murders across all of the states, or 15.4.

In a similar vein, we can estimate India's assault rate per 100,000 inhabitants to be 161. (approximated to the nearest integer value).

Since we don't have a source for India's urban population, we'll treat these the same way we did with any other missing figures.

Please take note that some of the assumptions we made prior to using this external dataset were for our own benefit.

Data Transformation:

As is well known, the data transformation process includes one or more of the following steps: normalization, summarization, noise removal, smoothing, and noise removal from data.

We shall use smoothing in this example because it is easier than summarizing and normalizing.

As we can see, Georgia has an abnormally high murder rate per resident according to our statistics, whereas New Hampshire has an exceptionally low murder rate per inhabitant. The likelihood is that these are anomalies. In this instance, we'll change the value of Murder for Georgia from 17.4 to 7.4. Similar to this, we will substitute 2.01 with 2.1 for Norway.

For Assault:

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

```
glimpse(dataset)
```

Data Reduction:

Data reduction aims to produce a reduced representation of the dataset that may be utilized to produce the same or comparable analytical results is what data reduction aims to produce.

With only 50 rows in the sample for our case, it is fairly modest. Imagine that we now have values for each of the 196 nations in the world as well as the geographical values for which the attribute values are specified.

In that situation, there are a lot of rows, so rounding up the murder rate for inhabitants to two decimal places could make more sense given the limited processing and storage resources you have available.

Such a massive dataset will require a significant quantity of storage space for each additional decimal place for every data point.

For assault:

```
dataset$Assault<-as.numeric(format(round(dataset$Assault, 0)))
```

For murder:

```
dataset$Murder<-as.numeric(format(round(dataset$Murder, 0)))
```

Data Discretization:

All of the attributes included in our dataset are continuous type, as can be seen (values in real numbers). The attribute values may need to be discretized into binary or categorical categories, though, depending on the model you wish to create.

For instance, you might want to categorize the murder rate for each resident into four groups: less than or equal to 1.00 per resident (represented by 0), more than 1.00 but less than or equal to 2.00 per resident (represented by 1), more than 2.00 but less than or equal to 5.00 per resident (represented by 2), and more than 5.00 per resident (represented by 3).

Here is a table that displays the data set:

| | Crime Type | Urban Population |
|-----|---------------|------------------|
| AB | Most Serious | Medium |
| AS | Most Serious | Small |
| AZ | Most Serious | Extra large |
| AK | Less serious | Medium |
| CF | Most Serious | Extra large |
| CR | Less serious | Extra large |
| CN | Least Serious | Extra large |
| DW | Less serious | Extra large |
| FL | Most Serious | Extra large |
| HW | Least Serious | Extra large |
| ID | Less serious | Medium |
| ILL | Most Serious | Extra large |
| IND | Less serious | Large |
| KS | Least Serious | Large |
| KT | Less serious | Medium |
| LS | Most Serious | Large |
| MN | Least Serious | Medium |
| ML | Most Serious | Large |
| MC | Less serious | Extra large |
| MI | Most Serious | Extra large |
| MG | Least Serious | Large |
| MS | Most Serious | Small |

| | | |
|-----|---------------|-------------|
| MI | Less serious | Large |
| MO | Least Serious | Medium |
| NB | Least Serious | Large |
| Nv | Most Serious | Extra large |
| NH | Least Serious | Medium |
| nk | Less serious | Extra large |
| NM | Most Serious | Large |
| NY | Most Serious | Small |
| NC | Most Serious | Small |
| No | Least Serious | Small |
| a | Less serious | Extra large |
| OK | Less serious | Large |
| OR | Less serious | Large |
| PS | Less serious | Extra large |
| RI | Less serious | Extra large |
| sC | Most Serious | Small |
| SD | Least Serious | Small |
| TS | Most Serious | Medium |
| TX | Most Serious | Extra large |
| UT | Less serious | Extra large |
| VM | Least Serious | Small |
| vG | Less serious | Large |
| w | Less serious | Extra large |
| WV | Least Serious | Small |
| VVC | Least Serious | Large |
| WY | Less serious | Medium |

Discussion and Conclusion:

In order to offer the necessary dataset, pre-processing techniques are used, and all missing data and outliers are found. without data pre-processing we can not get a perfect analysis result.