

**AMERICAN INTERNATIONAL UNIVERSITY BANGLADESH**

**Department of Computer Science**

**Project Title: Apply Data Pre-processing on a Dataset**

**Course:** **INTRODUCTION TO DATA SCIENCE MID, Fall:2022**

**Student Name: Sultanul Arifeen Hamim**

**Student ID: 20-42017-1**

**Section: B**

**Submitted To: AKINUL ISLAM JONY**

**Project Title: Apply Data Pre-processing on a Dataset**

**Project Overview:**

The main purpose of doing this project is to get a proper idea about Data preprocessing. This project mainly focuses on the data pre-processing which is a very important part of Data science. Course. Moreover, different types of pre-processing techniques are used here which also serve a purpose to get a good grip of them. so, one of its purposes can be defined by it. This project mainly focuses dataset which 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. Data pre-processing is a technique that uses a variety of approaches to clean up raw data and turn it into usable information. The information can then be used for a variety of tasks, including reporting and decision-making. In this project, some of the data are missing or incomplete, some of the data are noisy or contain outliers, and some of the data are inconsistent, which indicates that the data is not stored in accordance with the column's restrictions. With completing this project, it will increase the knowledge about data pre-processing using R language.

**Project solution design**:

Getting the dataset ready for data analysis I must first carry out the data pre-processing. I tidy up the data first. Analyzing a raw dataset to detect and remove duplicates, mistakes, and extraneous data is known as data cleaning. The table contained some missing data, which I filled in using median after replacing with N/A. Then I made an effort to manage every piece of noisy data that was present in the dataset. Following that, I was unable to locate any data Munging. The data set was relatively clean when the data cleaning procedure was completed. I then attempt to integrate the data by adding a new column with the name Type. So that this dataset is more consistent, I then perform the data transformation & data reduction

|  |
| --- |
|  |

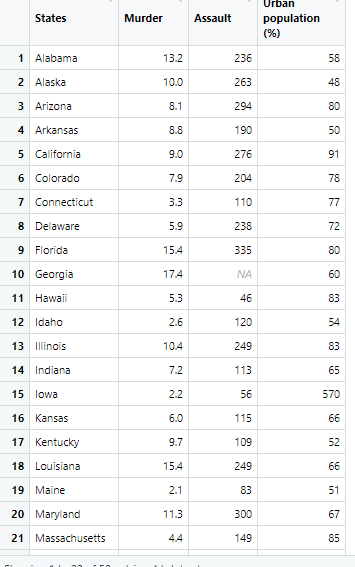
Flow-chart of Project solution design

**Data pre-processing:**

From pdf converted to excel then read the file using the following code:

library(readxl)  
dataset <- read\_excel("D:/AIUB 9th Semester/data science/dataset.xlsx")

Then view the data using the following code:  
View(dataset)

****

1. **Data cleaning 🡺**

**(a). Handling Missing Data:** In the Assault column For Georgia, the value is missing

dataset1<-dataset

dataset1$Assault<-dataset$Assault[is.na(dataset1$Assault)] =median (dataset1$Assault, na.rm = TRUE)

[So, here 1st ignore field with N/A and get median & finds field with N/A value and places median of assault]

|  |  |
| --- | --- |
| Before Handling Missing Data | After Handling Missing Data |

**(b). Smooth Noisy Data:** I think the Murder column has noisy data cause the people’s body count can’t be decimal. Outlier in urban population (%): Lowa 570, New York 6 & also in assault: South Carolina 879. So, it is replaced by using following code

dataset1$`Urban population (%) `[dataset1$`Urban population (%)` == 570]<-57

dataset1$`Urban population (%) `[dataset1$`Urban population (%)` == 6]<-60

dataset1$Assault [dataset1$Assault == 879] <-79

|  |  |
| --- | --- |
| Before Smooth Noisy Data | After Smooth Noisy Data |

**(c). Data Munging:** Here this process is not needed

1. **Data** **Integration🡺**Here Iprepare the dataset to integrate a new column (named type) based on the urban population variable. Converting the urban population percentage into types, for example, small (small (<30%), medium (<55%), large (<75%), and extra-large (75% and above) by using case\_when function allows you to vectorise multiple if\_else()

library(dplyr)

#Importing the dplyr library to use the group\_by() and summarise() functions

dataset1 <- dataset1 %>% mutate (Type = case\_when

(

(dataset1$`Urban population (%) `) < 30 ~ "small",

(dataset1$`Urban population (%) `) < 55 ~ "medium",

(dataset1$`Urban population (%) `) < 75 ~ "large",

(dataset1$`Urban population (%) `) >= 75 ~ "extra-large"

))

|  |  |
| --- | --- |
| Before Data Integration | After Data Integration |

1. **Data** **Transformation🡺** As Transformation suggested our data must be transformed so it is consistent and readable. So, in this project I use Normalization by decimal scaling processes for data transformation

dataset1$Assault = as.numeric(format(round(dataset1$Assault,0)))

dataset1$Type <- factor (dataset1$Type, ordered = TRUE)

|  |  |
| --- | --- |
| Before Data Transformation | After Data Transformation |

1. **Data** **Reduction 🡺**

In the murder section, I perform data reduction by ceiling the value in that section because murder cannot happen in the decimal or float number. As we know, data reduction aims to produce a reduced dataset representation that can be used to obtain the same or similar analytical results.

dataset1$Murder<-ceiling(dataset1$Murder)

# the output will be intergal

|  |  |
| --- | --- |
| Before Data Reduction | After Data Reduction |

1. **Data** **Discretization**🡺 I think in this data set we don’t need to apply the discretization.

|  |
| --- |
|  |

**Cleaned dataset**

**Discussion and Conclusion:**

This project was about a dataset that contains statistics in arrests per 100,000 residents for assault and murder, in each of the 50 US states, in 1973 population living in urban areas, where I have applied the data preprocessing process as Data preprocessing is the concept of changing the raw data into a clean data set. So, after completing the preprocessing I got a total clean data set to use, which involves data cleaning, data Integration, data transformation, data reduction & data discretization. After completing the project, I gained vast knowledge about. When data is gathered from different sources, it is collected in raw format, which is not feasible for the analysis, so here we need to convert the raw data into a clean data set by using data preprocessing. Here I have also additionally integrated values from this clean dataset into a raw dataset, which shows the types of categories of states based on their urban population. So, I personally believe that the project has been a success. Lastly, it can be said that this project’s main outcome is to increase the knowledge about data pre-processing using the R programming language.