

**AMERICAN INTERNATIONAL UNIVERSITY–BANGLADESH (AIUB)**

**FACULTY OF SCIENCE & TECHNOLOGY**

**DEPARTMENT OF COMPUTER SCIENCE**

**INTRODUCTION TO DATA SCIENCESummer 2024-2025**

**Section: A**

**Group: 17**

PROJECT REPORT ON

Applying Data Pre-processing on a Dataset

**Supervised By**

**TOHEDUL ISLAM**

**Submitted By**

|  |  |
| --- | --- |
| Name | ID |
| 1. MD. IMRAN AHMED | **20-43738-2** |
| 2. BORTOMAN DAS | **21-44429-1** |
| 3. MD. MIRAZUL HASAN | **22-46674-1** |

Date of Submission: **July 15, 2024**

**Introduction:**The dataset contains detailed information about caesarian section results for 80 pregnant women, highlighting significant characteristics of delivery problems in the medical field. Given the diverse data sources, the dataset may contain errors and inconsistencies, necessitating a thorough pre-processing stage to clean and transform the raw data into a format suitable for machine learning models. This pre-processing is crucial to ensure the data's accuracy and usability, enabling meaningful insights into the factors influencing caesarian sections and contributing to the improvement of maternal healthcare services in rural Bangladesh.

We classify delivery time to 0 = Timely, 1 = Premature and 2 = Latecomer. As for the blood pressure in three statuses of 0 = Low, 1 = Normal and 2 = High moods. Heart Problem is classified as 0 = apt and 1 = inept. And Caesarian as 0 = No and 1 = Yes.

**Data Exploration:**

**Import the dataset:**

The dataset is saved as the ‘MPD\_Sec-A.csv’ file. To begin preprocessing data in R Studio, we must import the file first. After importing the dataset, the data is stored in the ‘mydata’ variable,



A screenshot of a medical report

Description automatically generated

Figure 01: Imported main dataset.

**Required Library:**

For our data Pre-processing purpose, we will use ‘’, ‘tibble’, ‘dplyr’ and ‘ggplot2’ library,

A screen shot of a computer code

Description automatically generated

**Summary of the dataset:**

A screenshot of a computer screen

Description automatically generated

In the summary of the main dataset, we can see the dataset has 9 attributes where some attributes have missing values.

To show the data types of the attributes in the dataset,

A screenshot of a computer code

Description automatically generated

We can see there are 83 instances in the dataset. The ‘Patient\_id’, ‘Age’, ‘Delivery\_time’, ‘Heart’ and ‘Caesarian’ attributes are integer type, the ‘weight.kg’. attribute is numerical type and the ‘Gender’, ‘Delivery\_number’ and ‘Blood’ attributes are character type.

The dataset has 13 missing values distributed in the whole dataset,



**Project Solution Design:**

The dataset shows there are missing values in a couple attributes, there may be some invalid values. These values must be recovered properly. Moreover, if there are any outliers in particular attribute, we must recover them using proper method. Finally, we also need to convert attributes to their proper data type. For example, we need to convert the ‘Delivery\_time’, ‘Heart’ and ‘Caesarian’ attribute to character datatype. For recovering missing values, invalid values, or outliers we may use average (mean) or most frequent value (mode) depending on the datatype of the attribute.

**Data Pre-processing:**

**Patient\_id Attribute:**

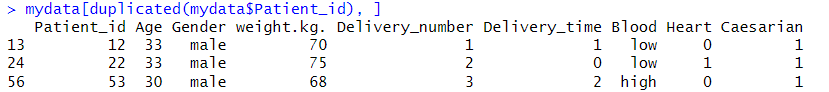
The ‘Patient\_id’ column is an integer type attribute.



There are no missing values in the ‘Patient\_id’ column.



There are 3 duplicated ids in the ‘Patient\_id’ column,



Remove the duplicate values from the ‘Patient\_id’ column,



*A screenshot of a computer

Description automatically generated*

Figure 02: Dataset after removing duplicate values from ‘Patient\_id’ column.

The column name is in the wrong format,

****

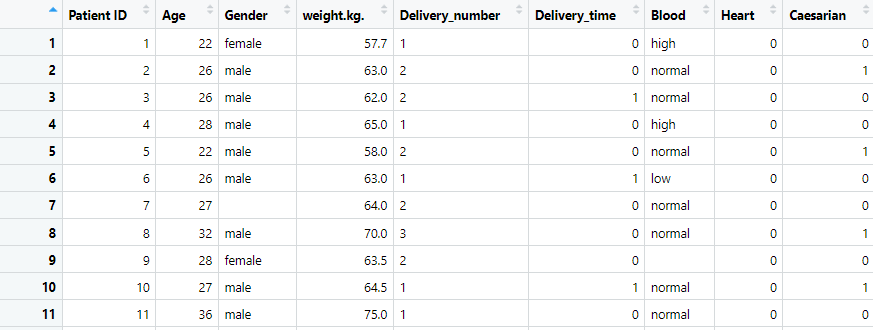
****

Figure 03: Dataset after renaming the ‘Patient\_id’ column to ‘Patient ID’.

**Age Attribute:**

The ‘Age’ column is an integer type attribute.

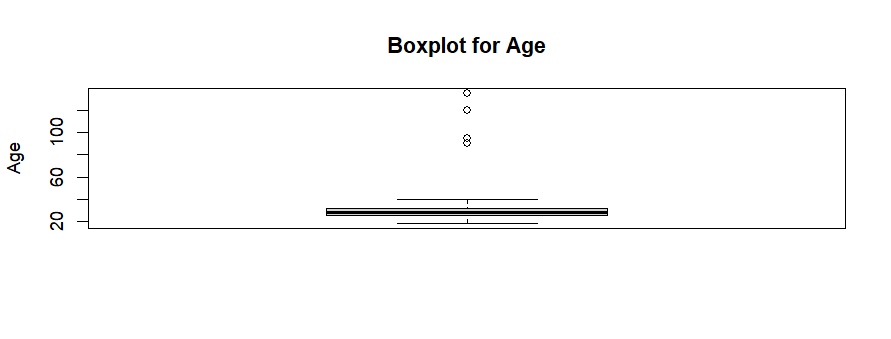


There are 4 missing values in the ‘Age’ column,



Check for outliers in the ‘Age’ column using boxplot,





There are 4 outliers in the ‘Age’ column,



Before recovering missing values, we need to remove outliers. Because we are going to use the average (mean) value to recover the missing values. If we didn’t remove the outliers the average (mean) values would not be accurate.

To remove the outliers, we will use Interquartile Range (IQR) method,

A white background with blue text

Description automatically generated

The outliers are removed, now we can apply the average (mean) values to recover the missing values.



A screenshot of a chart

Description automatically generated

Figure 04: Dataset after removing outliers & recovering missing values from ‘Age’ column.

**Gender Attribute:**

The ‘Gender’ column is a character type attribute,



Check what values are there in the ‘Gender’ column,



Here, the valid values are male and female but the value ‘mmale’, ‘Male-ish’, ‘maile’ and missing values are invalid. As the ‘Gender’ column is categorical, we will convert it in numerical values to recover any missing values or invalid values.

Replace values of ‘Gender’ column from with 1 and 2,



A screenshot of a chart

Description automatically generated

Figure 05: Dataset after replacing values of ‘Gender’ column.

There are 6 missing values in the ‘Gender’ column,



The ‘Gender’ column is categorical, recover these missing values using the most frequent (mode) value,



A screenshot of a computer

Description automatically generated

Figure 06: Dataset after recovering missing values using the most frequent value in the ‘Gender’ column.

Replace the ‘Gender’ attribute’s values back to ‘male’ and ‘female’.



A screenshot of a chart

Description automatically generated

Figure 07: Dataset after replacing the ‘Gender’ column values.

**weight.kg. Attribute:**

The ‘weight.kg.’ column is a numeric type of attribute,

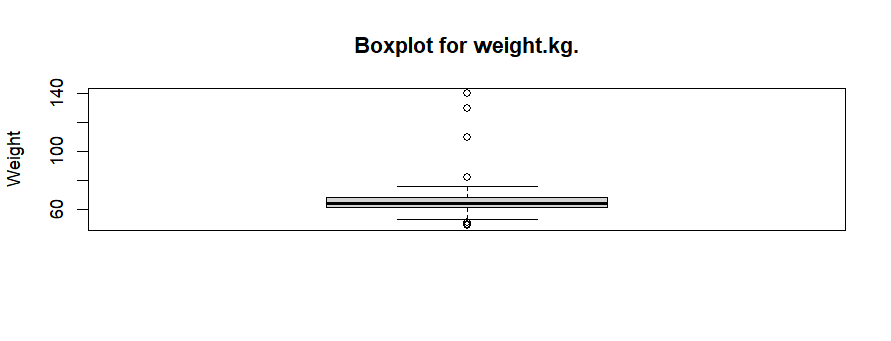


There are 4 missing values in the ‘weight.kg.’ column,



Check for outliers in the ‘weight.kg.’ column using boxplot,



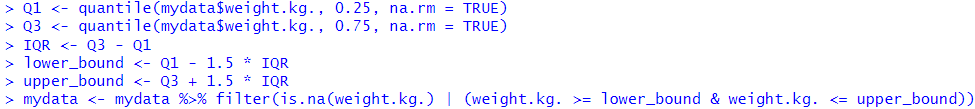


There are 7 outliers in the ‘weight.kg.’ column,



Before recovering missing values, we need to remove outliers. Because we are going to use the average (mean) value to recover the missing values. If we didn’t remove the outliers the average (mean) values would not be accurate.

To remove the outliers, we will use Interquartile Range (IQR) method,



The outliers are removed, now we can apply the average (mean) values to recover the missing values.



A screenshot of a computer

Description automatically generated

Figure 08: Dataset after removing outliers & recovering missing values from ‘weight.kg.’ column.

The column name is in the wrong format,



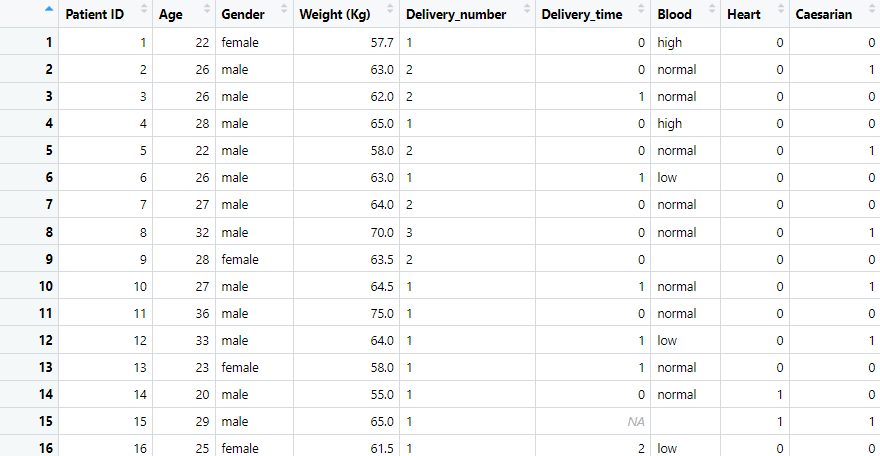


Figure 09: Dataset after renaming the ‘weight.kg.’ column to ‘Weight (Kg)’.

**Delivery\_number Attribute:**

The ‘Delivery\_number’ column is a character type of attribute,



Check what values are there in the ‘Delivery\_number’ column,



Here, the valid values are 1, 2, 3 and 4 but the value ‘1y’ and missing values are invalid. As the ‘Delivery\_number’ column is character type, we will convert it in numerical values to recover any missing values or invalid values.

A close up of a white background

Description automatically generated

After converting the ‘Delivery\_number’ column from character to integer type, all the invalid values are converted to missing values (NAs).

There are 2 missing values in the ‘Delivery\_number’ column,



Recover missing values using most frequent (mode) values,



A screenshot of a computer

Description automatically generated

Figure 10: Dataset after recovering missing values from ‘Delivery\_number’ column.

The column name is in the wrong format,



A screenshot of a data sheet

Description automatically generated

Figure 11: Dataset after renaming the ‘Delivery\_number’ column to ‘Delivery Number’.

**Delivery\_time Attribute:**

The ‘Delivery\_time’ column is an integer type of attribute,



Let’s check what values are there in the ‘Delivery\_time’ column,



Here, the valid values are 0, 1 and 2 but the missing values are invalid. As the ‘Delivery\_time’ column is integer type, we will recover any missing values and convert it to character type attribute.

There are 2 missing values in the ‘Delivery\_time’ column,



Recover missing values using most frequent (mode) values,



A screenshot of a data sheet

Description automatically generated

Figure 12: Dataset after recovering missing values from ‘Delivery\_time’ column.

We need to convert the ‘Delivery\_time’ column to character type attribute,



We will replace the values of ‘Delivery\_time’ column as like the description of the dataset

(0 = timely, 1 = premature, 2 = latecomer),



A screenshot of a computer

Description automatically generated

Figure 13: Dataset after replacing values ‘Delivery\_time’ column.

The column name is in the wrong format,

****

**A screenshot of a data

Description automatically generated**

Figure 14: Dataset after renaming the ‘Delivery\_time’ column to ‘Delivery Time’.

**Blood Attribute:**

The ‘Blood’ column is a character type of attribute,



Check what values are there in the ‘Blood’ column,



Here, the valid values are high, normal and low but the missing values are invalid. As the ‘Blood’ column is character type, we will convert it in numerical values to recover any missing values or invalid values,



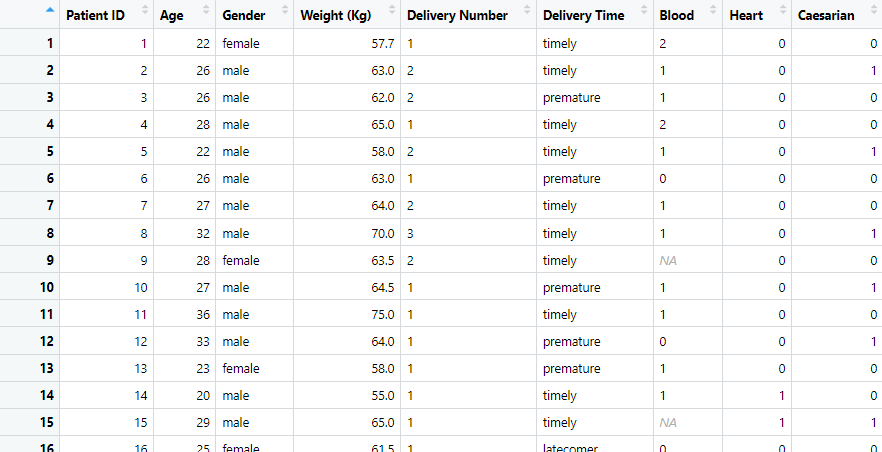


Figure 15: Dataset after replacing values of ‘Blood’ column.

There are 2 missing values in the ‘Blood’ column,



Recover missing values using most frequent (mode) values,



**A screenshot of a data

Description automatically generated**

Figure 16: Dataset after recovering missing values from the ‘Blood’ column.

Replace the ‘Blood’ attribute’s values back to the categorical value’s ‘low’, ‘normal’ and ‘high’,



A screenshot of a computer

Description automatically generated

Figure 17: Dataset after replacing the ‘Blood’ column values.

The column name is in the wrong format,



A screenshot of a computer

Description automatically generated

Figure 18: Dataset after renaming the ‘Blood’ column to ‘Blood of Pressure’.

**Blood Attribute:**

The ‘Heart’ column is an integer type of attribute,



The ‘Heart’ column has 0 and 1 as values and no missing or invalid values,



Convert the ‘Heart’ attribute’s values to the categorical value’s,



Replace the ‘Heart’ attribute’s values to ‘apt’ and ‘inept’,



A screenshot of a computer

Description automatically generated

Figure 19: Dataset after replacing the ‘Heart’ column values.

The column name is in the wrong format.



A screenshot of a computer

Description automatically generated

Figure 20: Dataset after renaming the ‘Heart’ column to ‘Heart Problem’.

**Caesarian Attribute:**

The ‘Caesarian’ column is an integer type of attribute,

A close up of a word

Description automatically generated

The ‘Caesarian’ column has 0 and 1 as values and some missing or invalid values,



There are 2 missing values,



Recover them using most frequent (mode) values,



Convert the ‘Caesarian’ attribute’s values to the categorical value’s,



Replace the ‘Caesarian’ attribute’s values to ‘No’ and ‘Yes’,



A screenshot of a computer

Description automatically generated

Figure 21: Dataset after replacing the ‘Caesarian’ column values.

After applying necessary pre-processing methods, the dataset looks like this,

A table of numbers and letters

Description automatically generated with medium confidence

Figure 22: Dataset after pre-processing.

From the summary we can see that all the attributes are in correct format and there are no missing or invalid values.

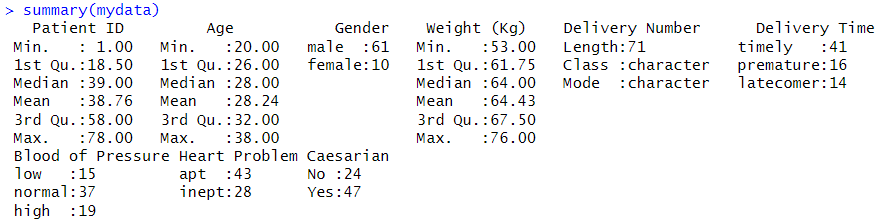


Figure 23: Dataset summary after pre-processing.

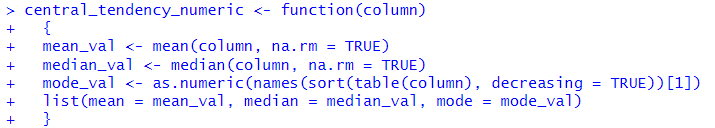
**Descriptive Statistics:**

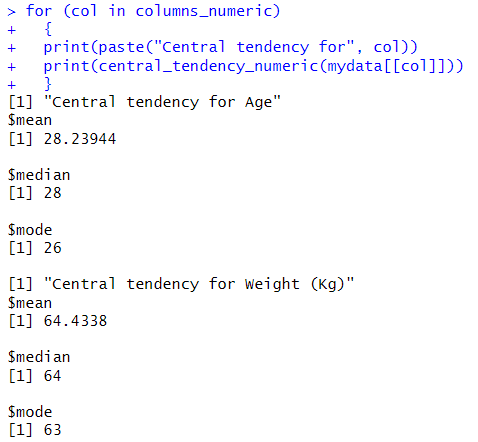
**Measure of Central Tendency:**

Basically ‘Central Tendency’ is measured by Mean, Median and Mode of continuous attributes,



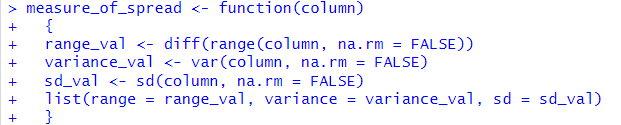






**Measure of Spread:**

Basically ‘Spread’ is measured by Range, Variance and Standard deviation of continuous attributes,



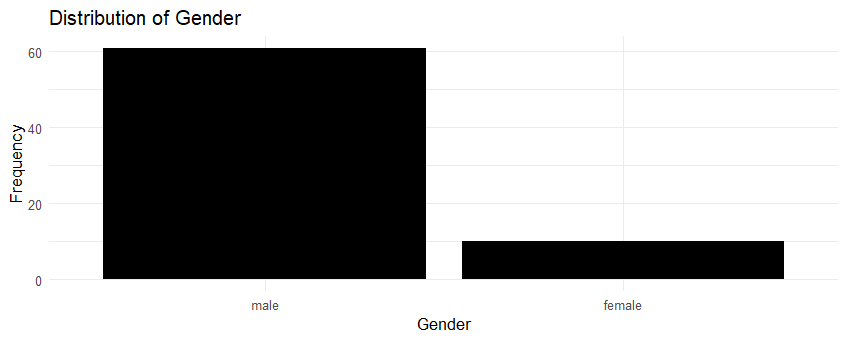
A screenshot of a computer code

Description automatically generated

**Distribution of Attributes Data:**

A graph of blue bars

Description automatically generated



A green bar graph with white background

Description automatically generated

A graph with purple squares

Description automatically generated

A graph of a number of orange rectangular objects

Description automatically generated with medium confidence

A yellow square with white text

Description automatically generated with medium confidence

A graph with red squares

Description automatically generated

A graph with blue squares

Description automatically generated