A picture containing calendar

Description automatically generated

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

**FACULTY OF SCIENCE AND TECHNOLOGY**

**MIDTERM ASSIGNMENT**

**INTRODUCTION TO DATA SCIENCE**

**Spring 2022-2023**

**Section: D**

**Submitted By**

**AL SHAKIB E ELAHI**

**ID: 20-43665-2**

**Department: CSE**

**Supervised By**

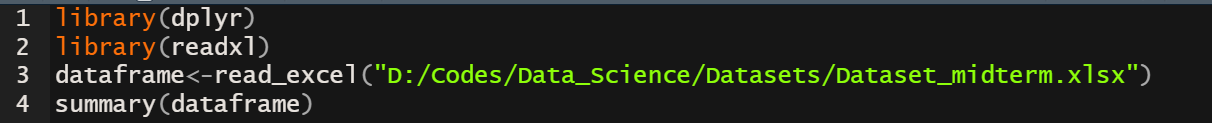
**Tohedul Islam**

**Assistant Professor**

**Department of Computer Science**

Date of Submission: **March 12, 2023**

**Data set import:**

****

**Output:**

**A picture containing graphical user interface

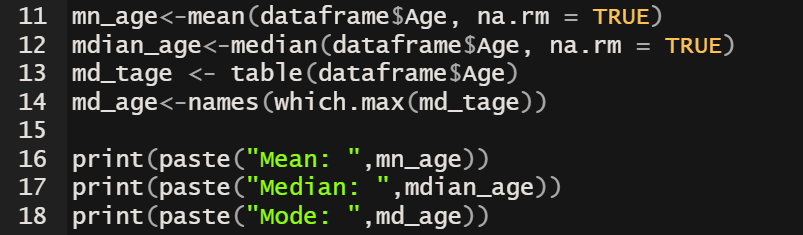
Description automatically generated**

**Data Pre-processing:**

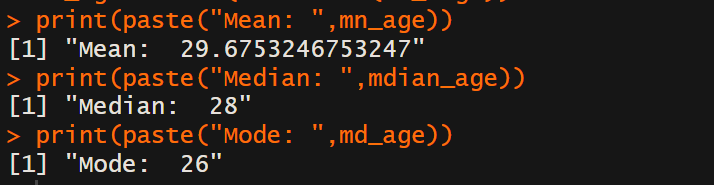
**Age Section**

From summary, it is observed that “Age” section contains 3 null values.

* **Detecting null values’ row number:**
* **Measure of Center Tendency:**

****

**Output:**

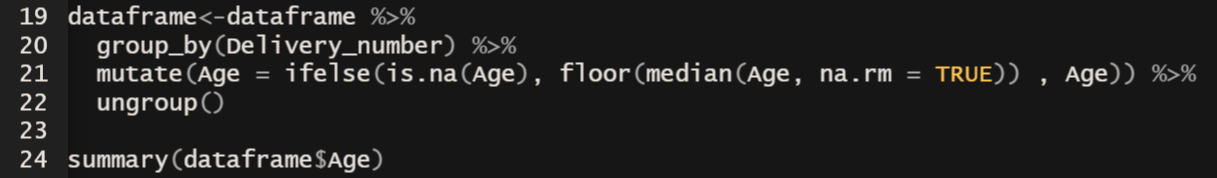
****

**Explanation:**

* na.rm == TRUE is used to skip the null (N/A) values
* table() provides the number of frequencies for a certain Age
* which.max() gives the max frequency value and associated Age
* names() extracts the maximum frequency
* paste() function mainly combines the string and numeric data type for the output.

Here mean is slightly greater than median and the mode.

* **Replacing null values with median group by Delivery\_number:**

****

**Output:**

**Graphical user interface, text

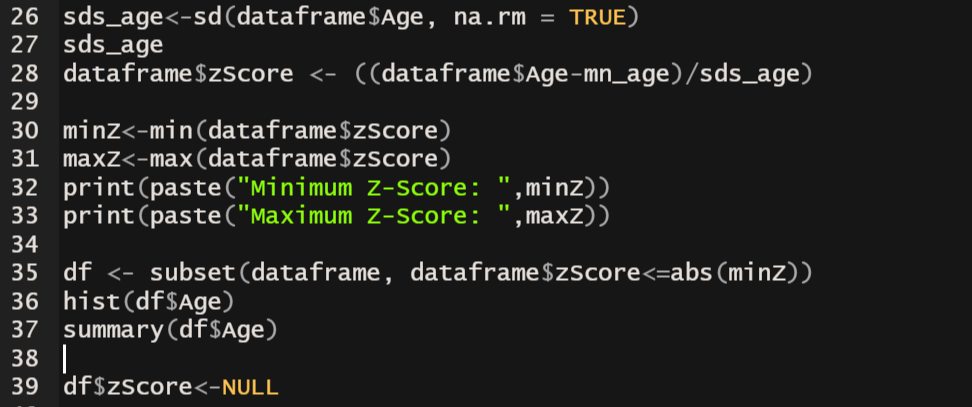
Description automatically generated**

**Explanation:**

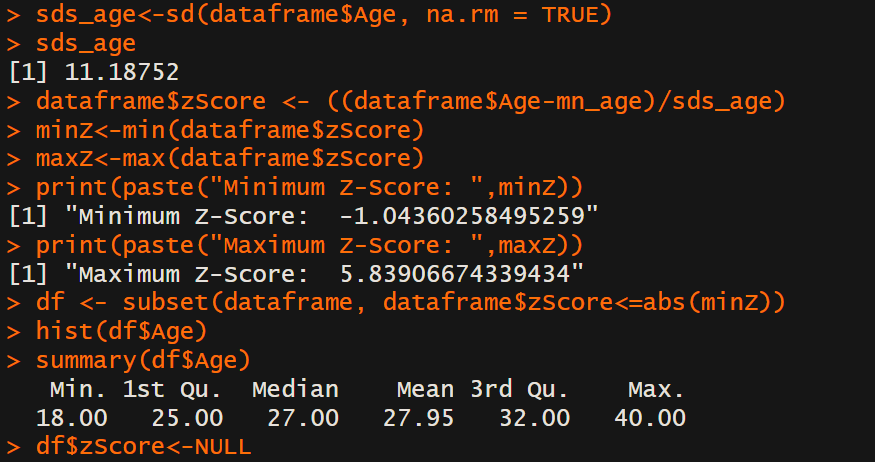
* group\_by() creates subsets according to each unique value of Delivery\_number
* ungroup() removes the group created by group\_by()
* ifelse() is to allow conditional operations on dataframe
* mutate() if for modifying existing dataframe information with the median

In my observation, it is obvious that Age is slightly dependent on number of Deliveries.

* **Outliers Detection:**



**Output:**

****

**Graph:**

Chart, histogram

Description automatically generated

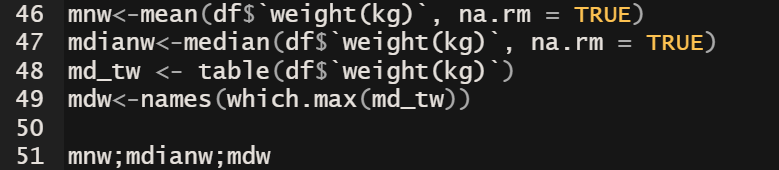
**Explanation:**

First, standard deviation of Age is calculated. Then Z-Score is determined. It is obvious that, for a normal distribution, maximum and minimum value for Z-Score should be almost similar. But in this case, maximum is far away and that’s mean, there is a outlier. So with subset() function, a subset of dataframe was taken with threshold of absolute value of the minimum value of the Z-Score. Lastly, to eliminate zScore column, it has been assigned NULL.

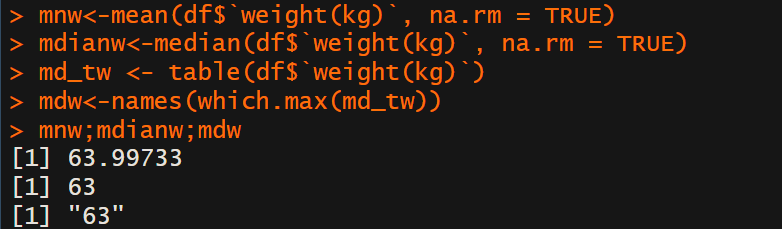
**Weight Section**

From summary, it is observed that “Weight” section also contains 3 null values.

* **Detecting null values’ row number:**
* **Measure of Center Tendency:**

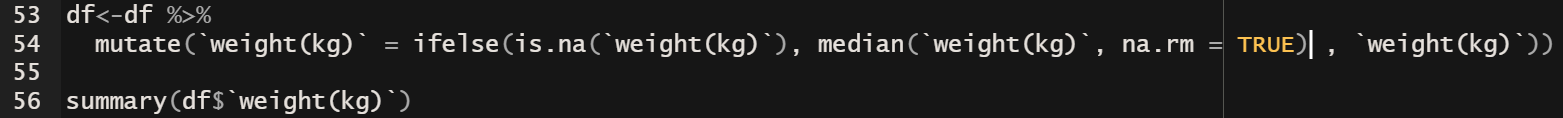
****

**Output:**

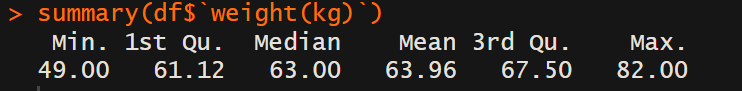
****

Here mean, median and mode are almost equal.

* **Replacing null values with median:**

****

**Output:**

****

**Explanation:**

* ifelse() is to allow conditional operations on dataframe
* mutate() if for modifying existing dataframe information with the median
* **Outliers Detection:**

Text

Description automatically generated

**Output:**

**Text

Description automatically generated**

**Graph:**

**Chart, histogram

Description automatically generated**

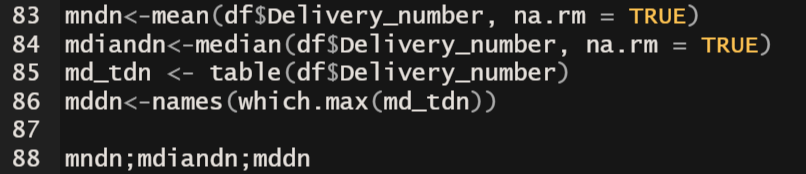
**Explanation:**

Here, difference between the maximum and minimum Z-Score is quite low and so there is no outliers according to my observation.

**Delivery Number**

From summary, it is observed that “Delivery\_number” section also contains 2 null values.

* **Detecting null values’ row number:**
* **Measure of Center Tendency:**

****

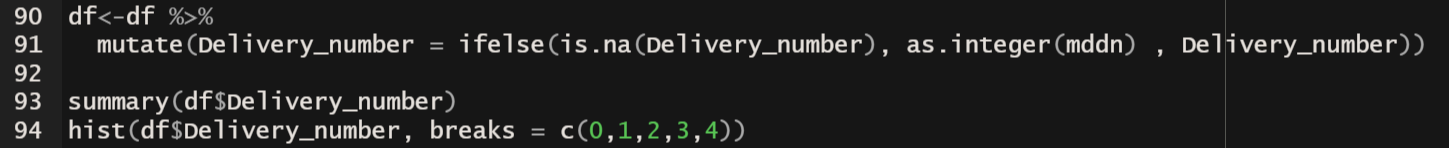
**Output:**

**Text

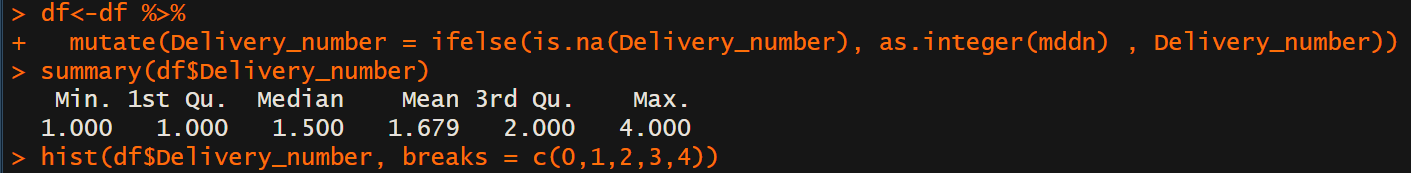
Description automatically generated**

Here mean, median and mode are in a certain range.

* **Replacing null values with mode:**

****

**Output:**

****

**Explanation:**

* as.interger() is used to change the data-type of “mddn” from character to integer as Delivery\_number is numeric.
* In hist() breaks is used to personalized.

**Graph:**

**Chart, histogram

Description automatically generated**

**Delivery Time**

* **Detecting null values’ row number:**

**Text

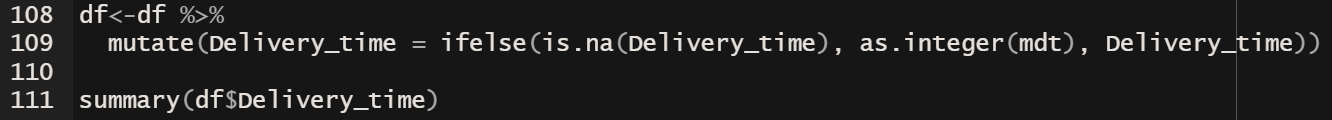
Description automatically generated**

* **Measure of Center Tendency:**

**Text

Description automatically generated**

* **Replacing null values with mode:**

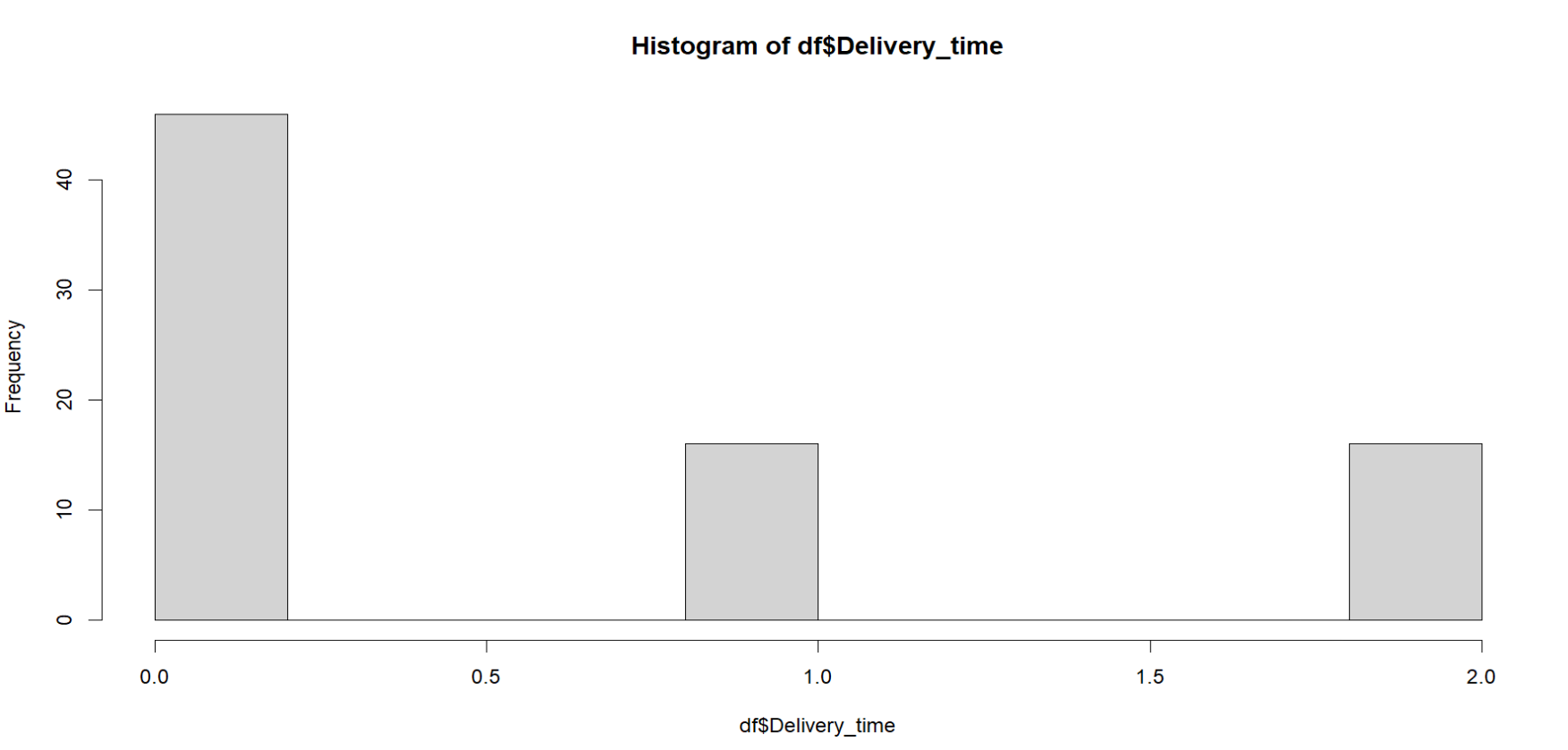


**Output:**

**Text

Description automatically generated**

**Graph:**

****

**Blood**

* **Detecting null values’ row number:**

Text

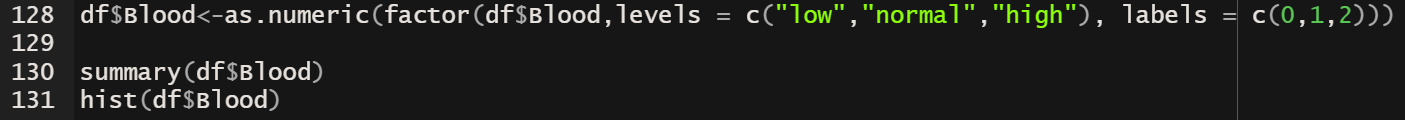
Description automatically generated with medium confidence

* **Replacing null values with mode:**

Text

Description automatically generated

* **Converting to numeric:**

****

**Graph:**

**Chart, waterfall chart

Description automatically generated**

**Ceasarian**

* **Omitting null values:**

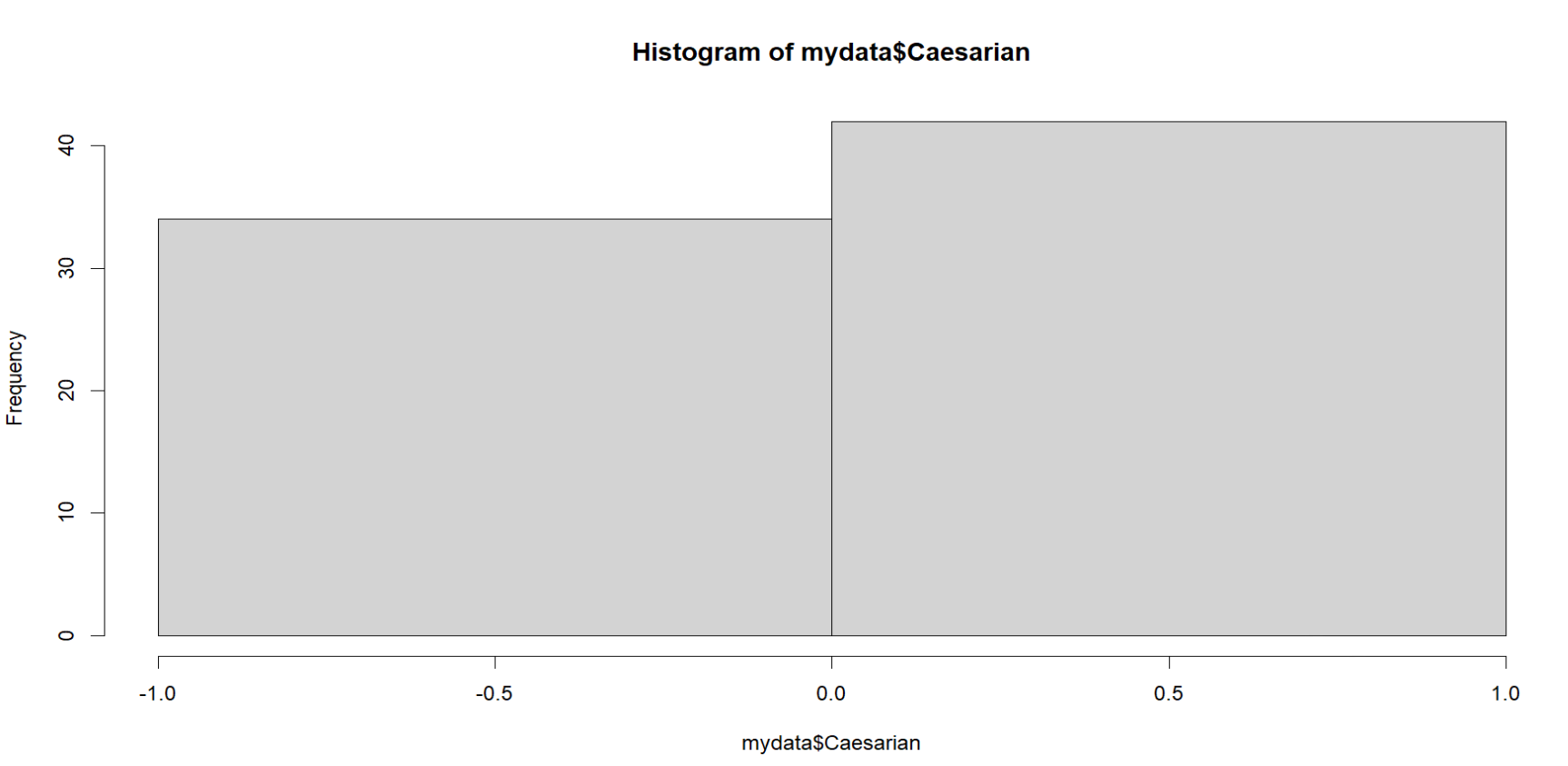
Text

Description automatically generated with low confidence

Explanation:

In my observation, caesarian is the dependent variable that is going to be predict and it is a sensitive data. So I preferred to omit this instead of replacing.

**Graph:**



* **Summary:**

**A picture containing graphical user interface

Description automatically generated**