**American International University – Bangladesh**

**Department of Computer Science & Engineering**



**Project Title:** **Apply data preparation steps (which can be applied) and do the univariate data exploration for the given data set.**

**Course: Introduction to Data Science**

|  |  |
| --- | --- |
| **Submitted by-** | **Submitted to-** |
| **Name: Pranay Acharjee**  **ID: 20-42372-1**  **Section: E**  **Spring 2022-2023**  **B.Sc. CSE** | **Name: Tohedul Islam** |

**Dataset Description:**

This is a Caesarian Delivery Prediction Dataset that contains information about 80 pregnant women who were classified based on their age, weight, delivery number, blood pressure, heart status, and delivery time. The dataset includes 8 attributes: Id, Age, Weight, Delivery Number, Delivery Time, Blood Pressure, Heart Problem, and caesarian. The target variable, Caesarian, indicates whether or not each woman had a caesarian delivery.

Attributes:

**Id:** The id of each woman.

**Age:** The age of the woman in years.

**Weights:** The weights of each woman.

**Delivery Number:** The number of previous deliveries the woman has had.

**Delivery Time:** The time of delivery for the current pregnancy, classified as Premature, Timely, or Latecomer.

**Blood Pressure:** The woman's blood pressure, is classified as Low, Normal, or High.

**Heart Problem:** Whether or not the woman has a heart problem, classified as Apt or Inept.

**Caesarian** (Target variable): Whether or not the woman had a caesarian delivery, classified as No or Yes.

Purpose: The Caesarian Delivery Prediction Dataset can be used to build a predictive model to identify whether a woman is likely to have a caesarian delivery based on her age, weight, delivery number, blood pressure, delivery time, and heart status.

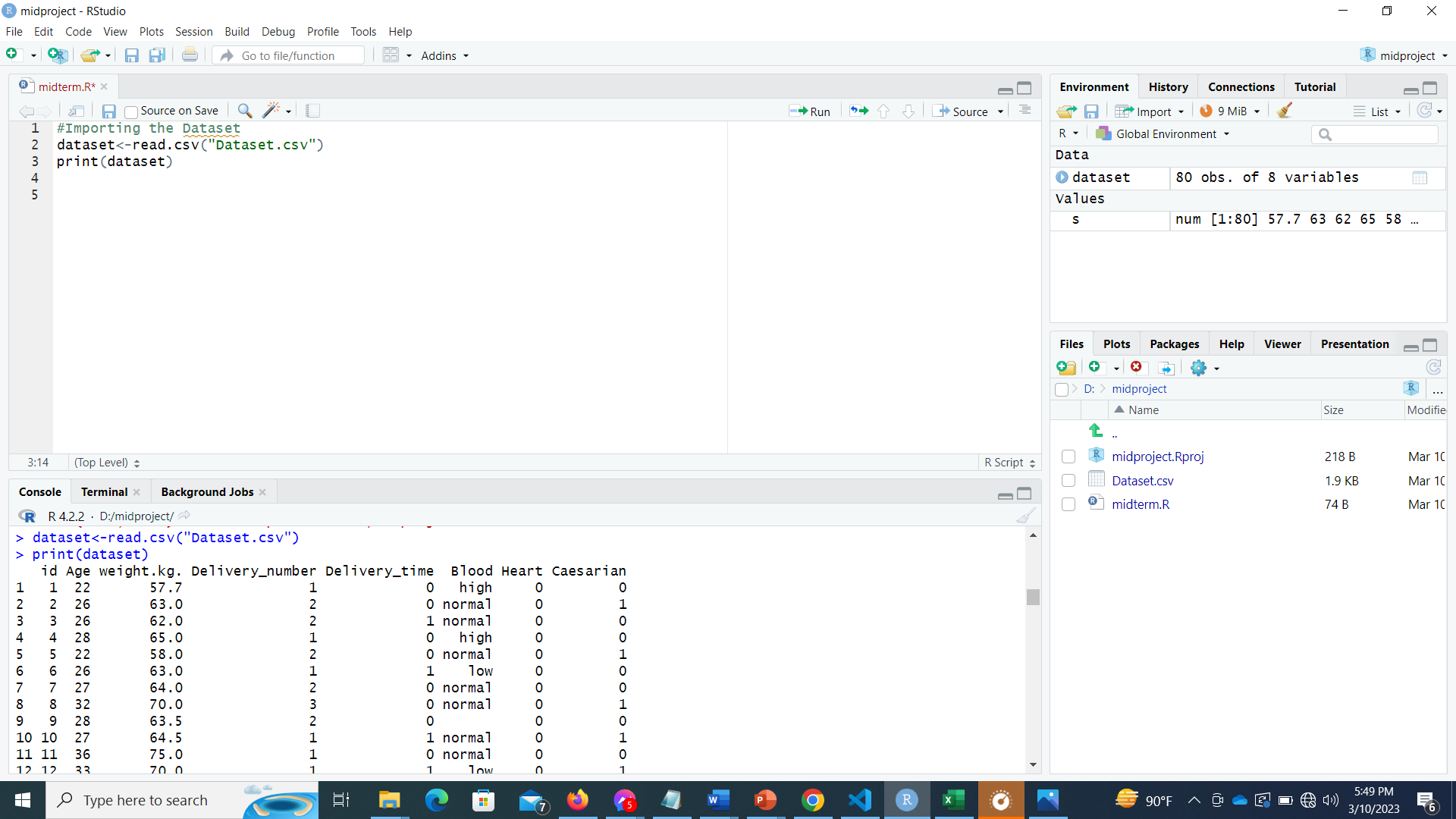
**Project Overview:**

Data pre-processing is a crucial stage in data analysis that involves converting raw data into a clear and comprehensible format that can be easily analysed by computers and machine learning systems. In practice, raw data can be messy and riddled with errors, requiring cleaning before it can be used for a specific purpose. Furthermore, we need to do univariate exploration, which involves analysing individual variables in a dataset separately, without considering the relationships between variables.

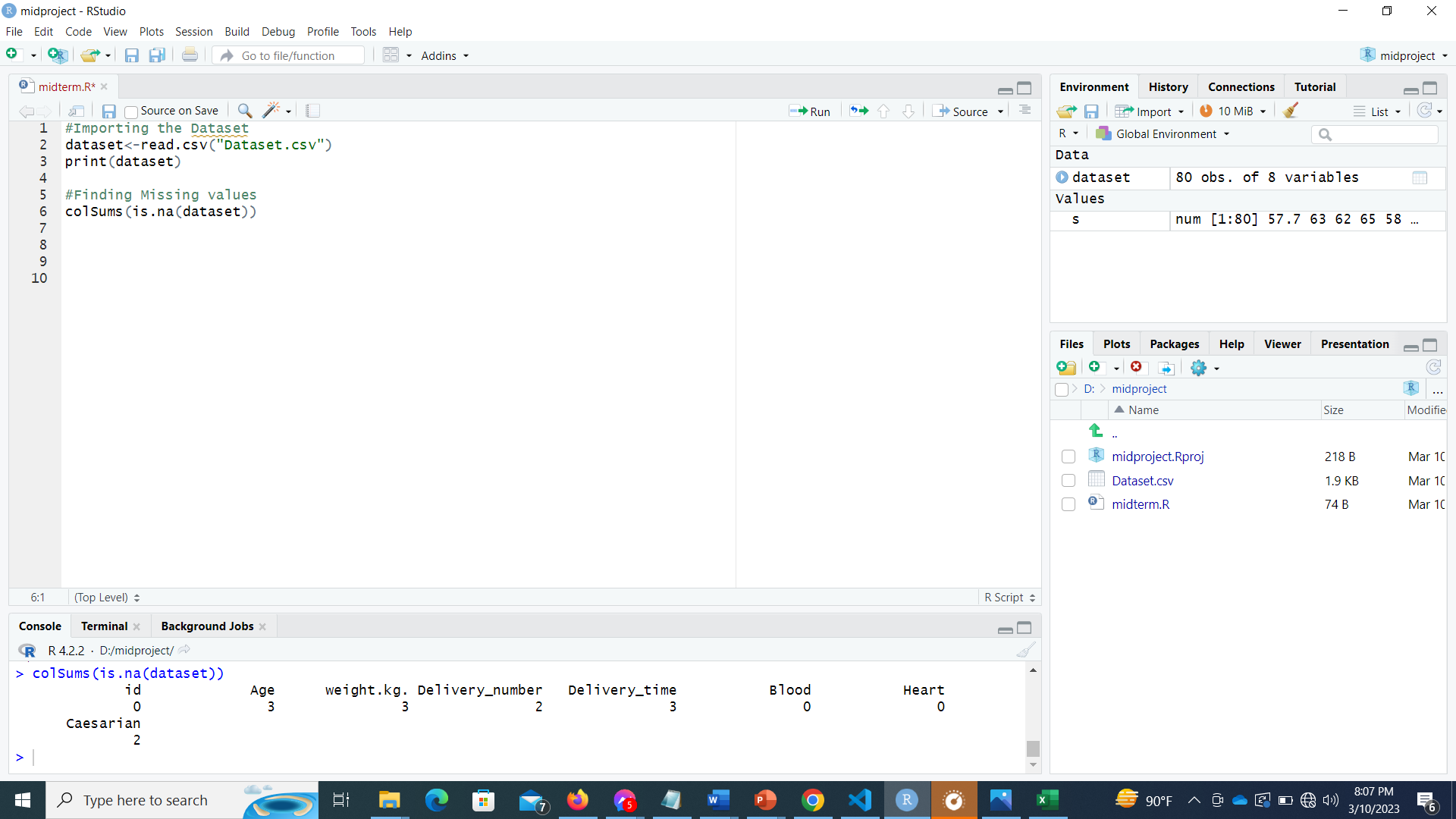
We can see that the dataset is not in a clean format. The dataset needs to be pre-processed and cleaned before any further use.

**Data pre-processing:**

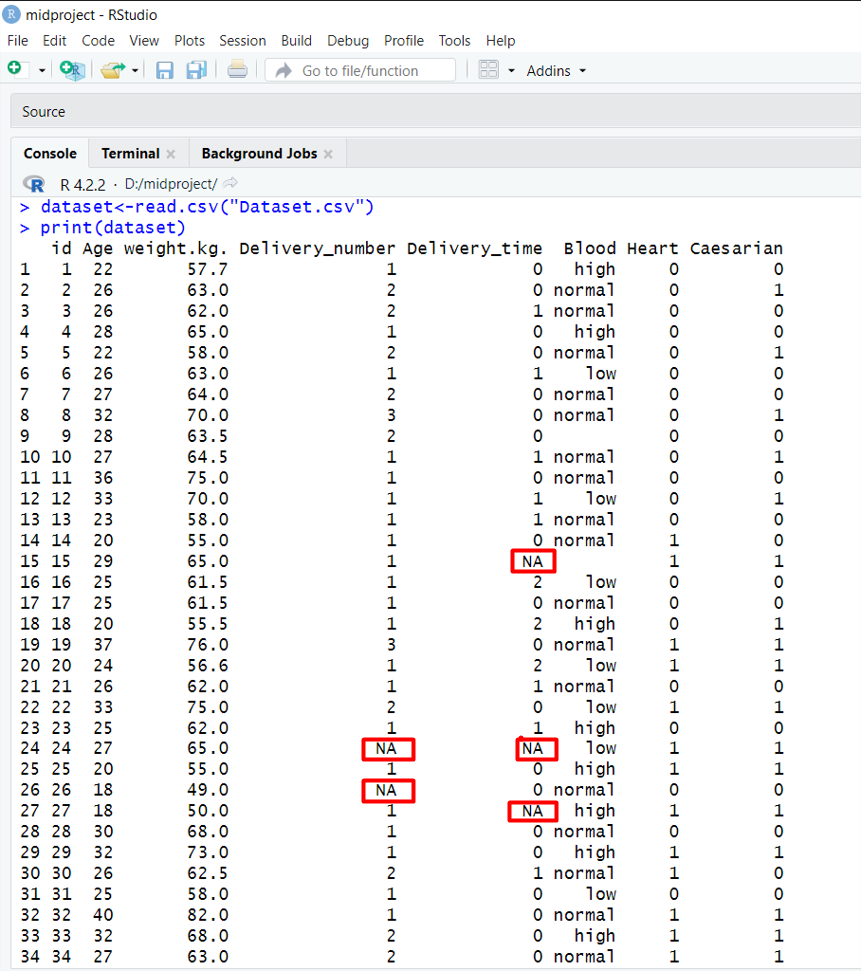
1. **Importing the Dataset:** The dataset is located in a file called dataset.csv in the current working directory. To begin data pre-processing using R, the first step is to import the dataset. Once imported, the dataset.csv file is transformed into an R data frame and stored in a variable named "dataset". After printing the dataset, it looks like this-

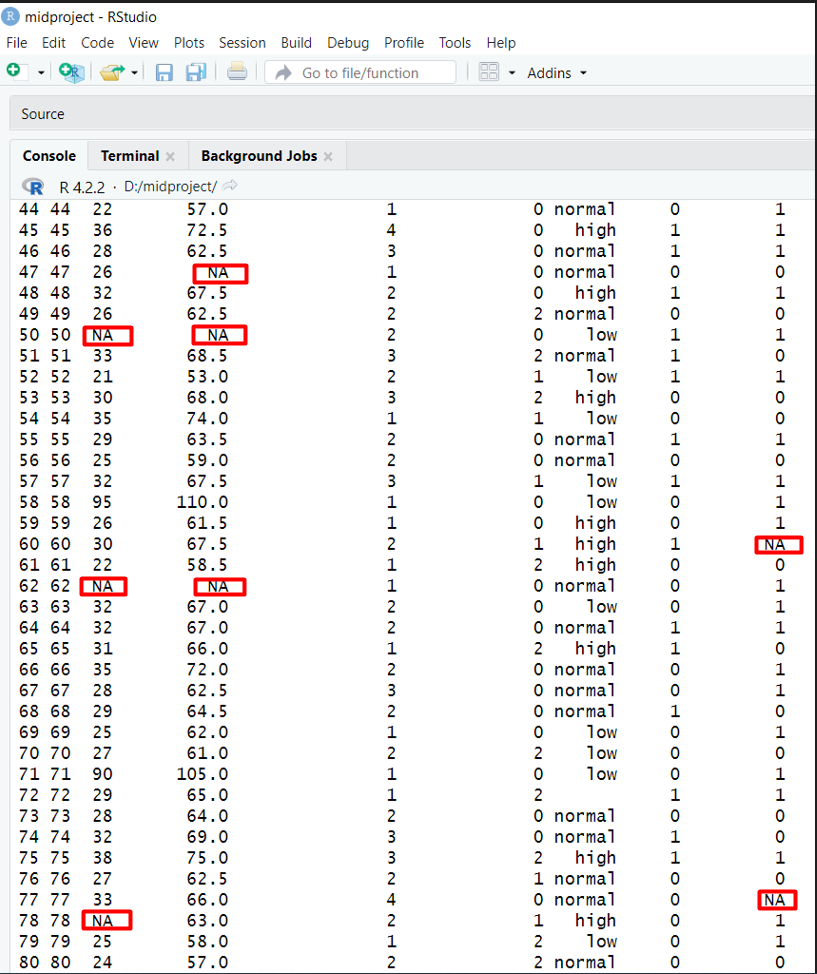


1. **Dealing with Missing Values:**
   1. We can see from the dataset that, there is some missing value (NA) present in column name- Age[3], Weight[3], Delivery Number[3], Delivery Time[2], and cesarean [2]. We can find out the missing values in this way:

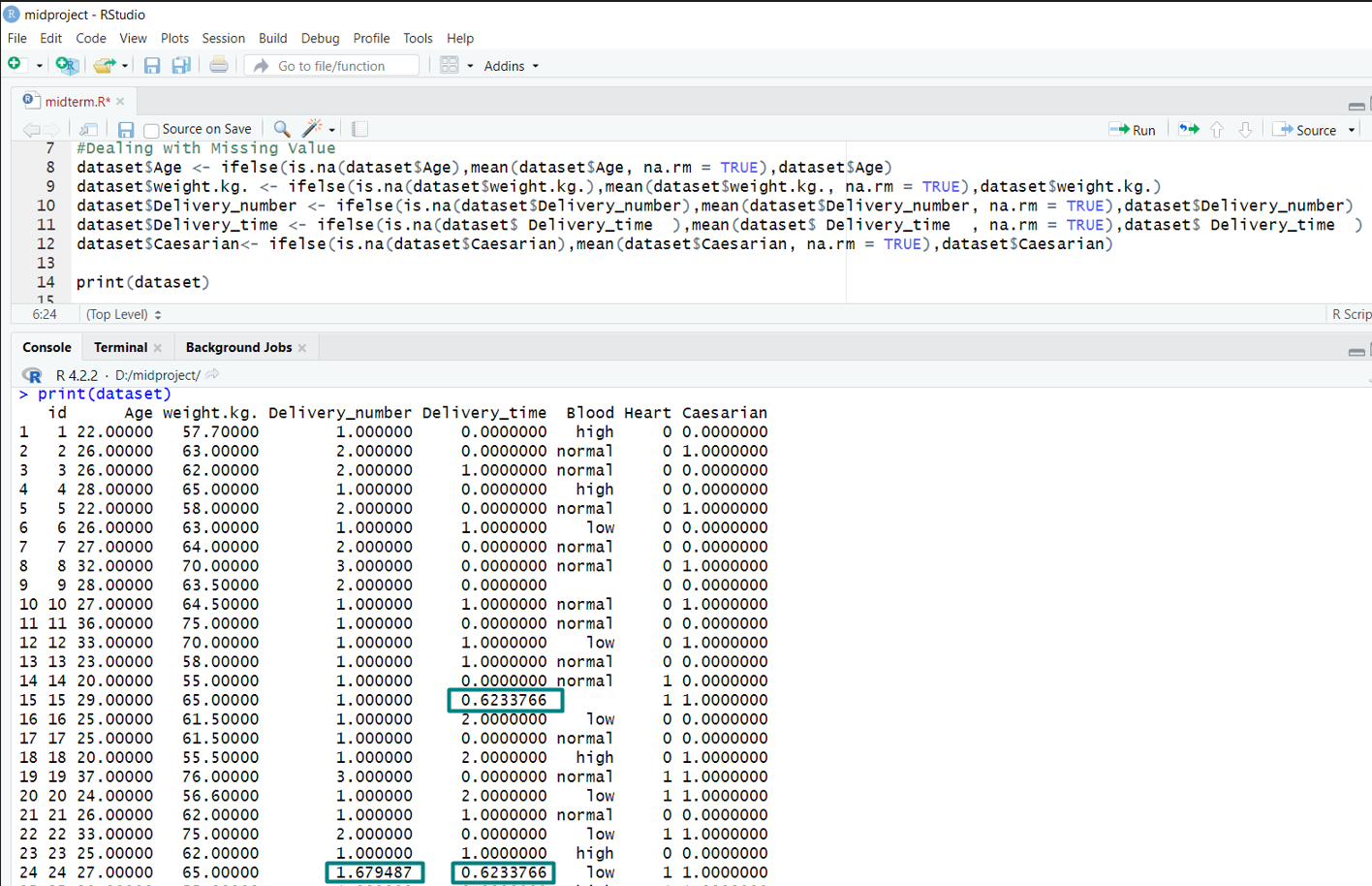


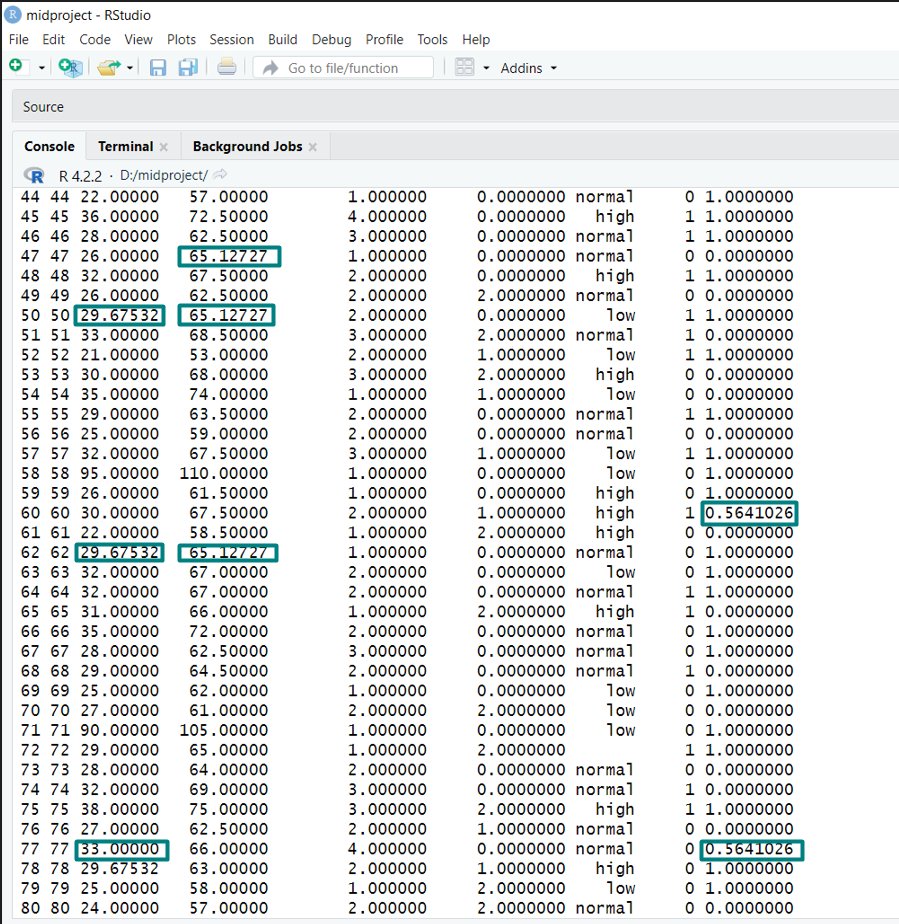
Before, the dataset looked like this-





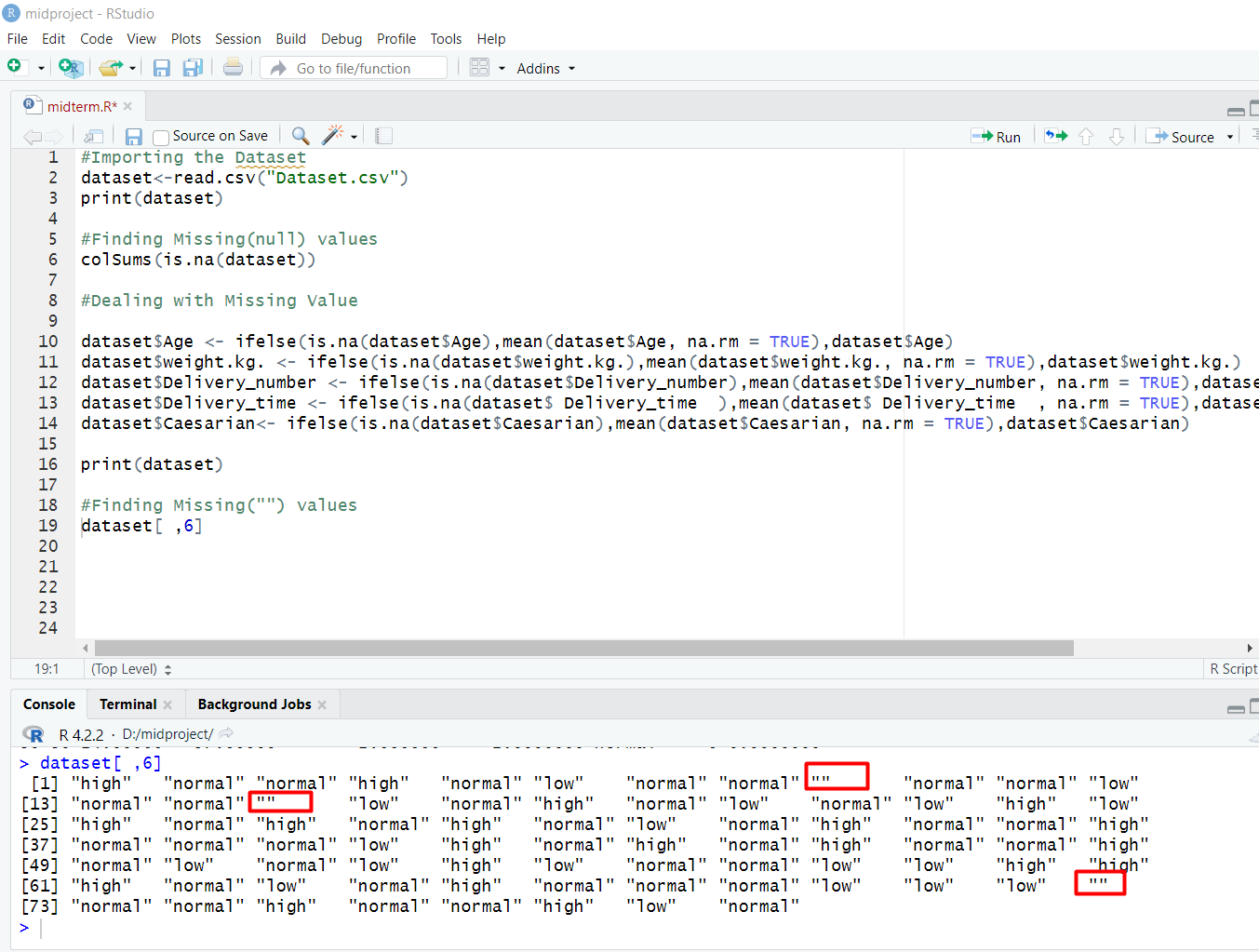
**2.2** Now, as these columns are in the numerical format we can replace the missing value with the mean value of those columns. R code for replacing missing value by the mean, Now, the dataset looks like this-



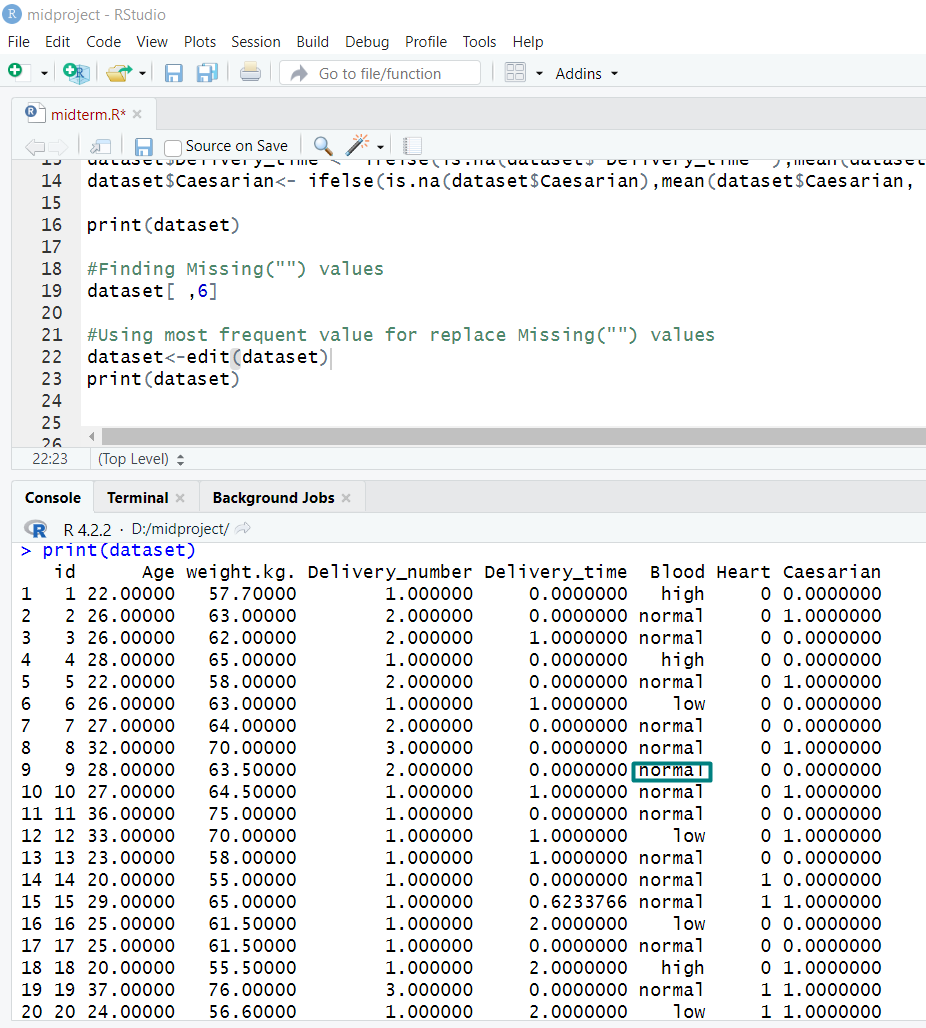


**2.3** Here we can see that in the “Blood” column, some values are missing.

We can find it out in this way-

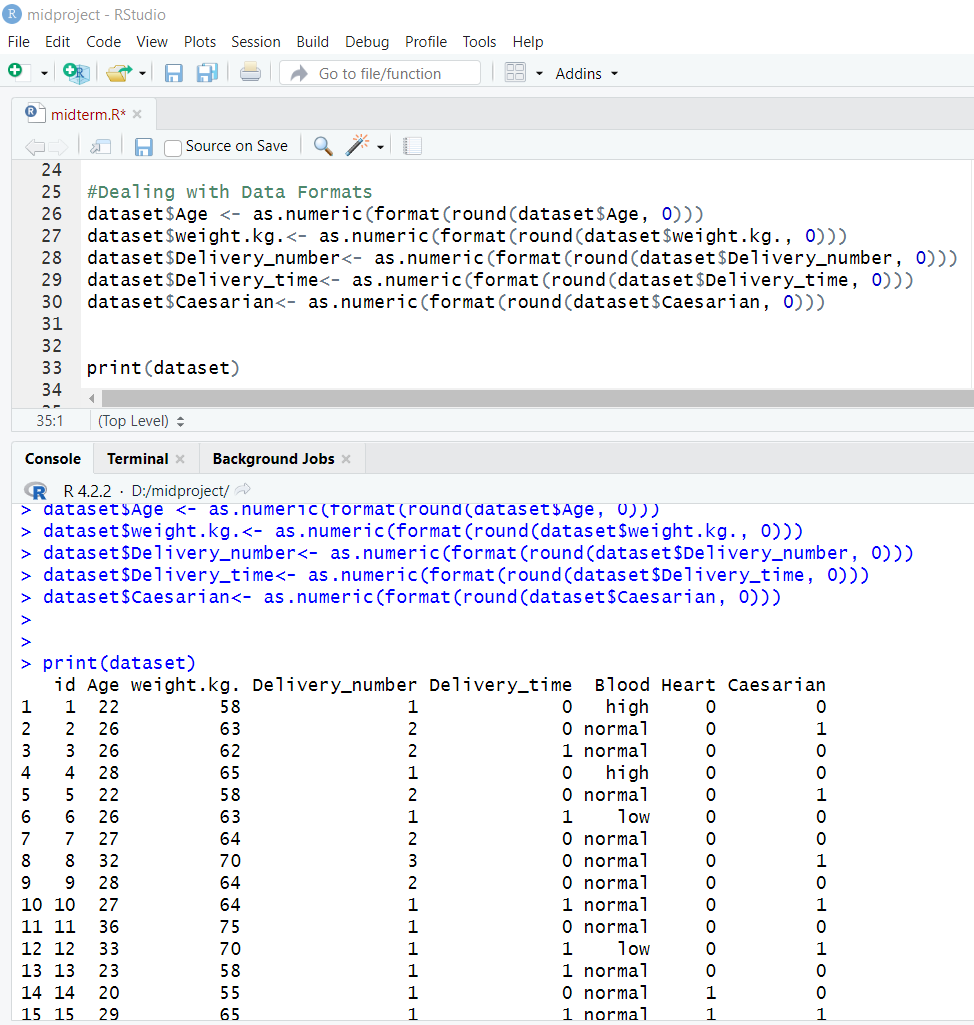


As the Blood column is categorical so we can overcome this problem using the most frequent value in the place of missing value.



1. **Dealing with Data types and Conversion:**

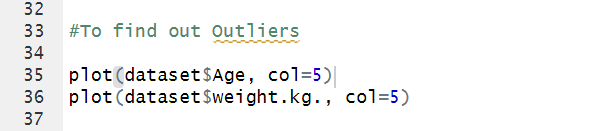
Now we can see that the maximum column contains decimal places in the data after dealing with null values in those columns. As we are not interested in having decimal places for those columns, we will round it up. We can round those variables in this way-



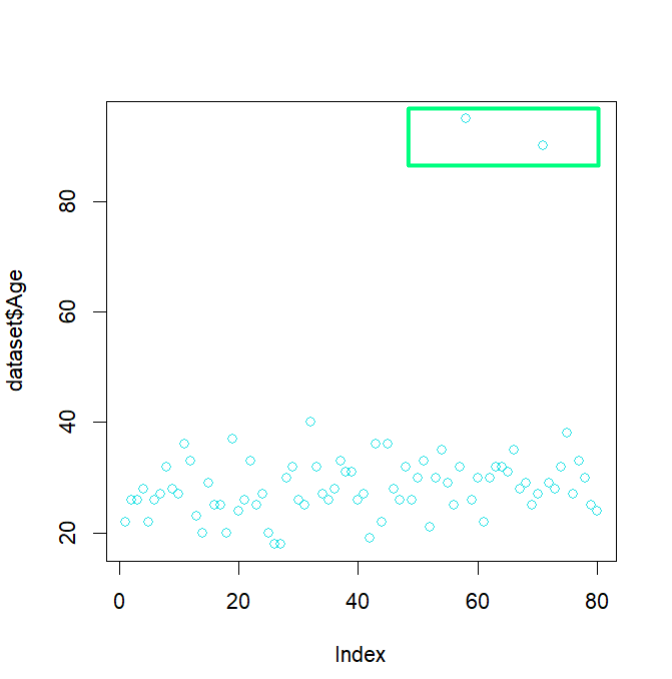
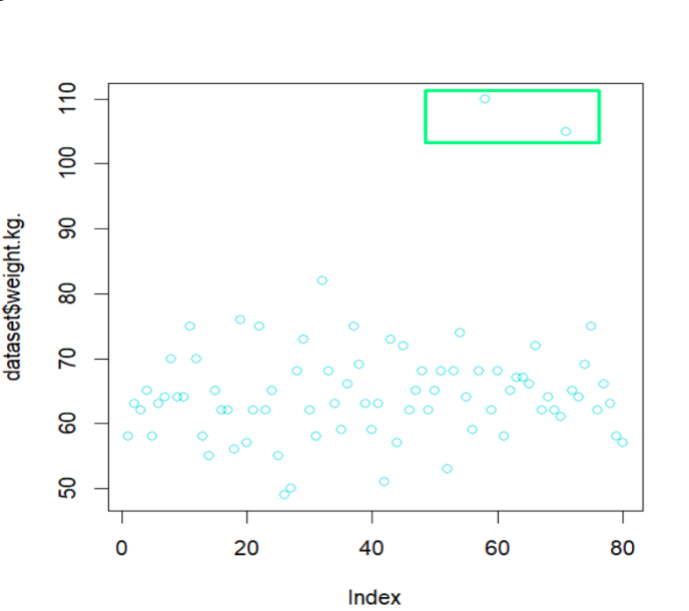
1. **Dealing with Outliers:**

Here, we can check whether there are any outliers or not. So to find it out we can use this way-

**R code-**



**Output-**

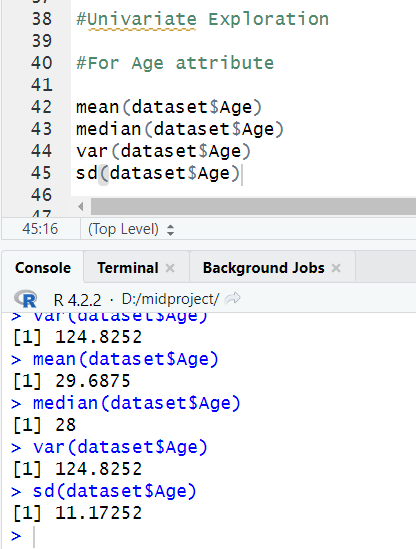


So, here we can see that in column $Age and $weight. kg., there are two points or data which are quite different from the rest of the points. So this can be called Outliers.

**Univariate data exploration:**

Univariate exploration in data science involves analyzing individual variables in a dataset one at a time, without considering the relationship between variables. This type of analysis is useful for gaining a basic understanding of the distribution, central tendency, and variability of a variable.

1. So, here we can find out the exploration of the Age attribute-

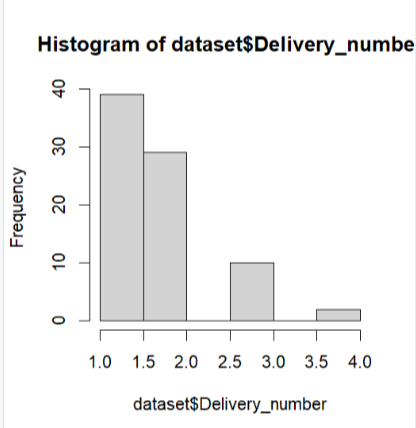


And here for weight.kg. attribute



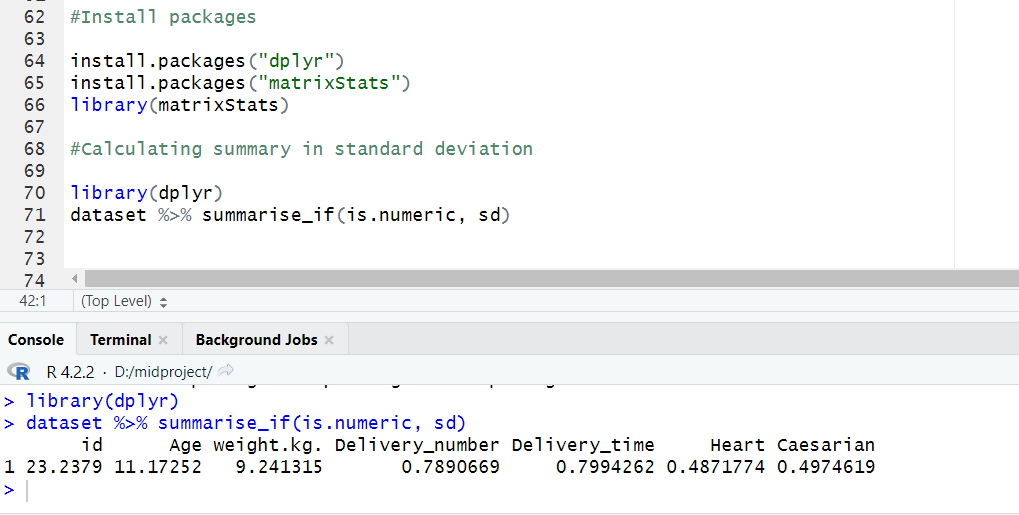
So, here we found the mean for the Age attribute 29.68 and the standard deviation 11.17. And for the weight.kg. attribute mean is 65.17 and the standard deviation is 9.24.

1. Now, we draw a histogram for Delivery\_number and Delivery\_time attributes for analysis.



From the histogram, we observe that most of the pregnant women were classified based on their Delivery number which means The number of previous deliveries the woman has had, these numbers are between one to two, and the pregnant women were classified based on their Delivery time which means The time of delivery for the current pregnancy, most of the case is premature.

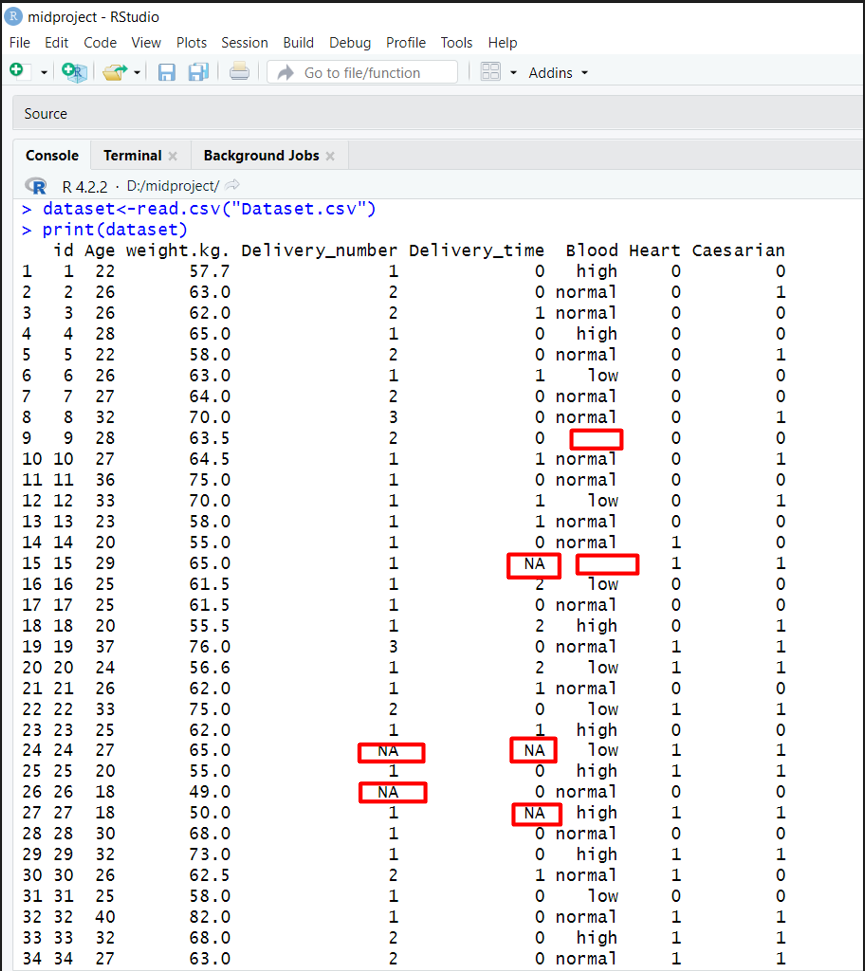
1. Standard deviation of each attribute



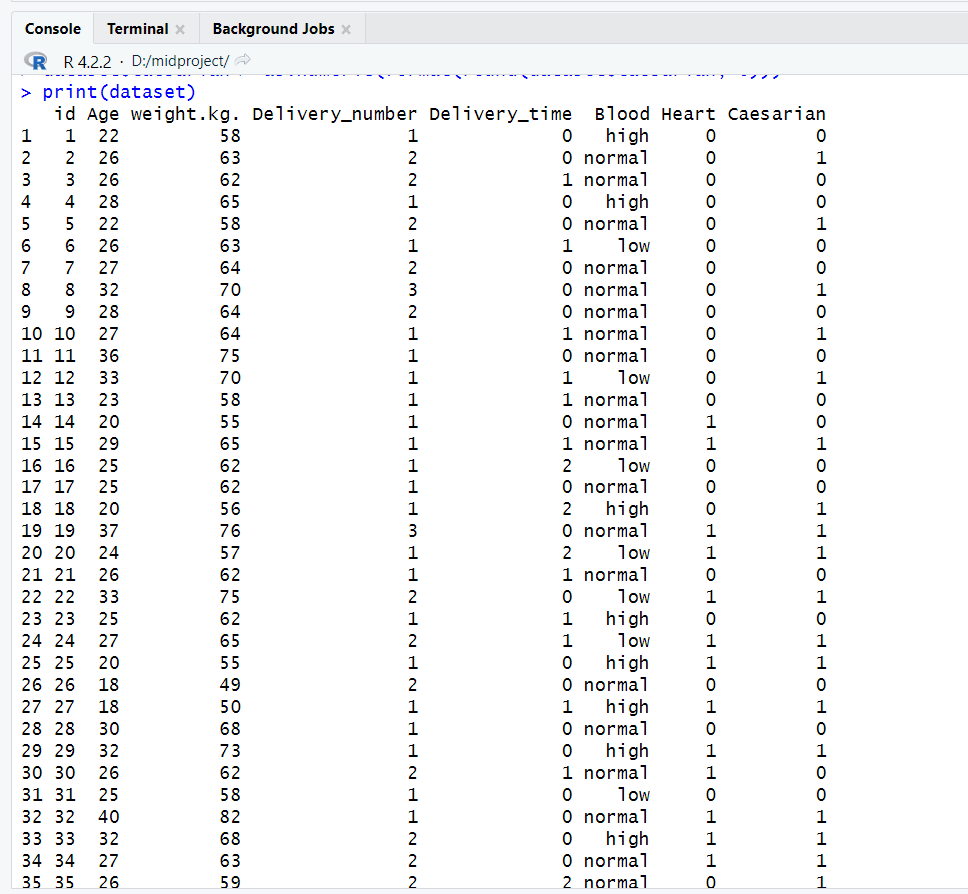
Here, we calculate the standard deviation of each numerical attribute.

**Discussion & Conclusion:**

At the beginning of the project, we were given a dataset that was totally messy. Null values, missing values, and outliers were present in this dataset. The dataset was like this-



After Applying data preparation steps and the univariate data exploration for the given data set., we got the dataset looks like this-



Now, we can use this clean, pre-processed dataset for further use.