A picture containing text, gambling house

Description automatically generated

**Department of Computer Science**

**American International University-Bangladesh**

**Mid Term Report**



 Course Name:  **INTRODUCTION TO DATA SCIENCE**

“A Report on Data Pre-Processing”

**Supervised By:**

**Tohedul Islam**

**Submitted By:**

**Sajjad Hossein**

ID: 20-43639-2

Section: [E]

.

**Project Title: Applying Data Pre-processing on a Dataset.**

**Project overview:**

Raw, real-world data such as text, I mages, and videos are messy. Not only can they contain errors and inconsistencies, they are often incomplete and lack a regular and consistent design. Machines like to process information in good order, reading data as 1's and 0's. As such, calculating structured data such as integers and percentages is straightforward. However, unstructured data in the form of text and images must first be cleaned and formatted before analysis. Data preprocessing refers to the steps that transform or encode data so that it can be easily interpreted by a computer. In order for the model to make accurate and accurate predictions, the algorithm must be able to quickly interpret the attributes of the data. Due to their diverse origins, most real-world datasets are particularly vulnerable to missing, inconsistent, and noisy data. Applying data mining algorithms to this noisy data yields poor results because they cannot recognize patterns. Therefore, data preprocessing is important for improving overall data quality. Data cleaning is a step in the data preprocessing process to fill missing values, smooth noisy data, fix discrepancies, and remove outliers. Data transformation includes data cleaning and data reduction techniques to transform data into an appropriate format. Data transformation includes data cleaning and data reduction techniques to transform data into an appropriate format. Data transformation is an important data preprocessing technique that must be performed on the data before data mining to provide easy-to-understand patterns. This is a systematic process of data preprocessing.

**Data Frame:**

Data Frames are tables with data displayed in them. Data Frames can contain various types of data. The first column can be character, while the second and third columns can be numeric or logical. Each column, however, should contain the same type of data.

**The Software and Language utilized for this project:**

The language we are going to use to conduct the project is R and the software we’re going to use to process data and shape data is R Studio. R language is designed specifically for statistical computing and analysis. R has powerful tools for data visualization, which helps to analyze and interpret complex data sets more easily. It allows us to create high-quality graphs, charts, and other visual representations of data. And R Studio is a powerful and easy way to interact with R programming. R Studio has built-in data visualization tools that allow users to create high-quality graphs, charts, and other visual representations of data.

**Importing Data from CSV File**

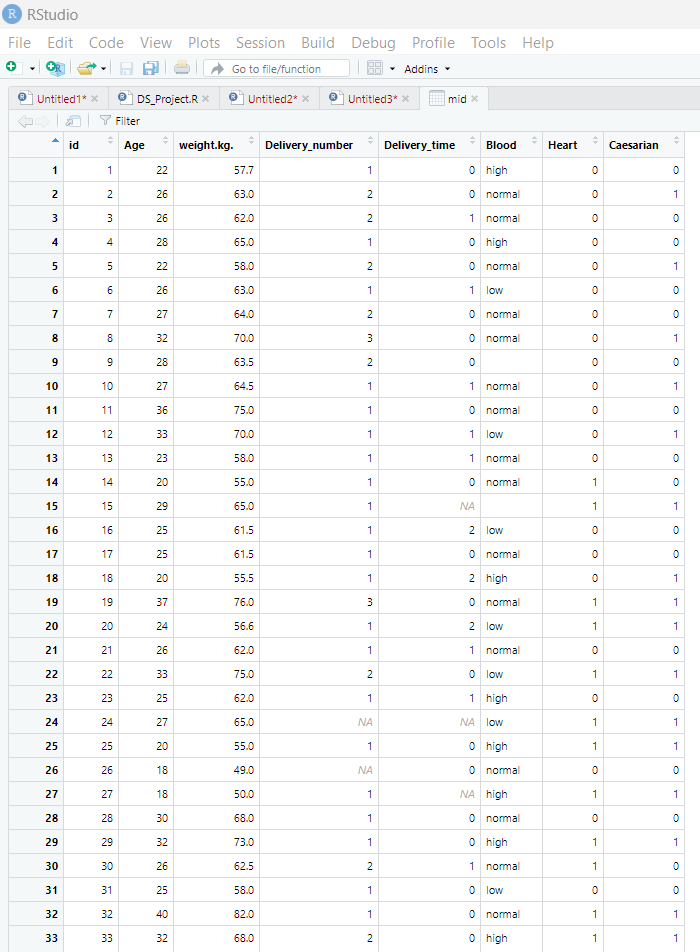
**CODE:**

# Calling CSV File

mid = read.csv("D:/Dataset\_midterm.csv",header = TRUE,sep = ",")

mid

mean(mid$Age)



**Data Pre-processing:**

**Data Munging:** When data is in the unstructured format, we perform data munging. Since in this dataset all the data are in structured format, there are no data munging steps needed in the data set as the data is in structured format.

**Handling Missing Data:** To deal with missing data, at first we must need to identify the missing values in the dataset.

When we tried to calculate the mean with the missing value in AgeColumn we will get:

mean(mid$Age)

[1] NA

**Finding Missing Value for Age (By Mean)**

**Code:**

#Missing value For Age

meanAge=mean(mid$Age,na.rm=TRUE)

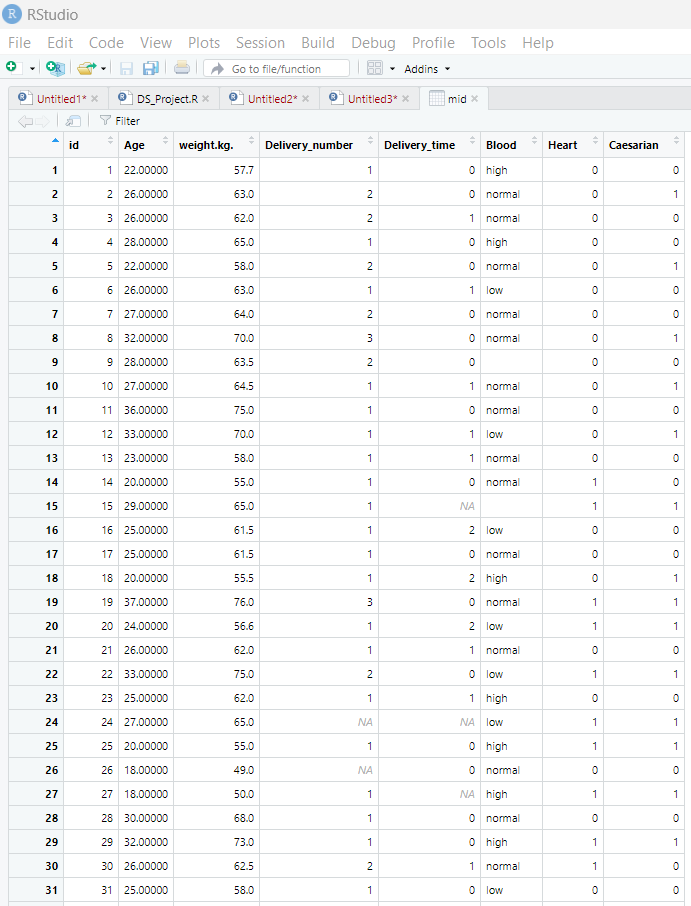
meanAge

**Replacing Missing Value for Age**

**Code:**

mid [is.na(mid$Age),"Age"] = meanAge

mid



Here Missing value for Age is placed.

**Finding & Replacing Missing Value for Age (By Median)**

**Code:**

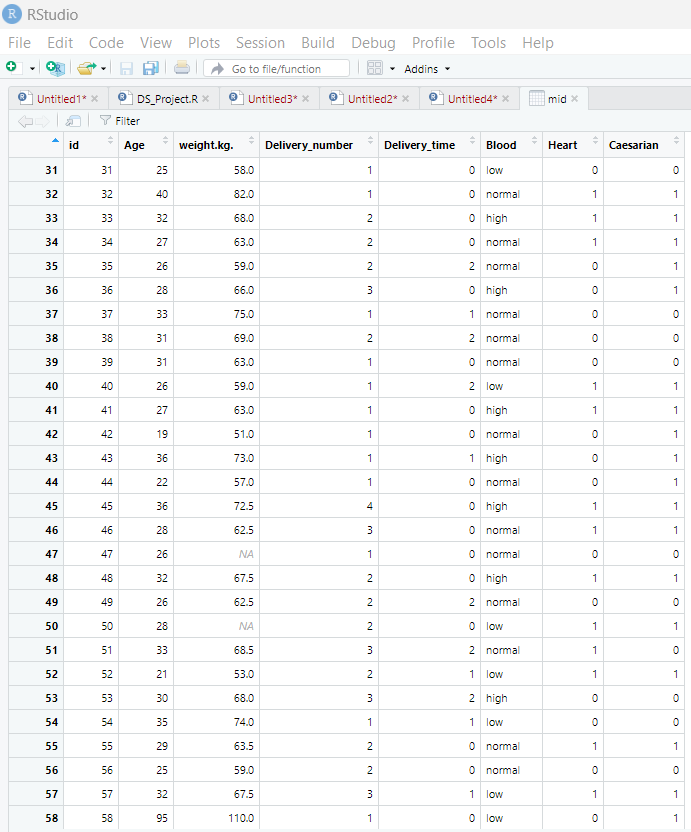
#median Age

medianAge=median(mid$Age,na.rm=TRUE)

medianAge

mid [is.na(mid$Age),"Age"] = medianAge

mid



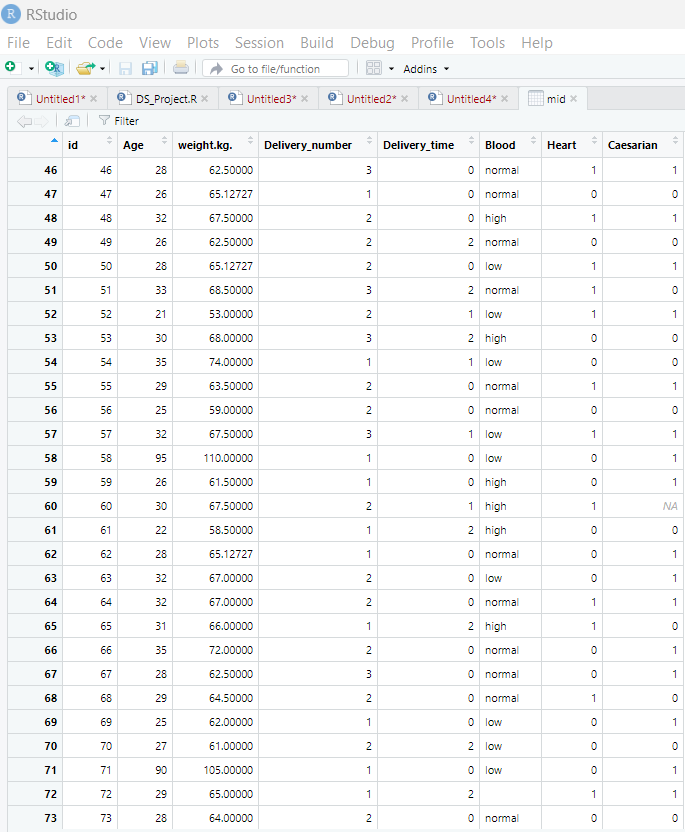
**Finding & Replacing Missing Value for Age (By Mode)**

**Code:**

#mode for age

library(DescTools)

mid$Age[is.na(mid$Age)] <- Mode(mid$Age,na.rm = TRUE)



**Finding Missing Value for weight.kg. (By Mean)**

**Code:**

#Missing value For weight.kg.

meanweight.kg.=mean(mid$weight.kg.,na.rm=TRUE)

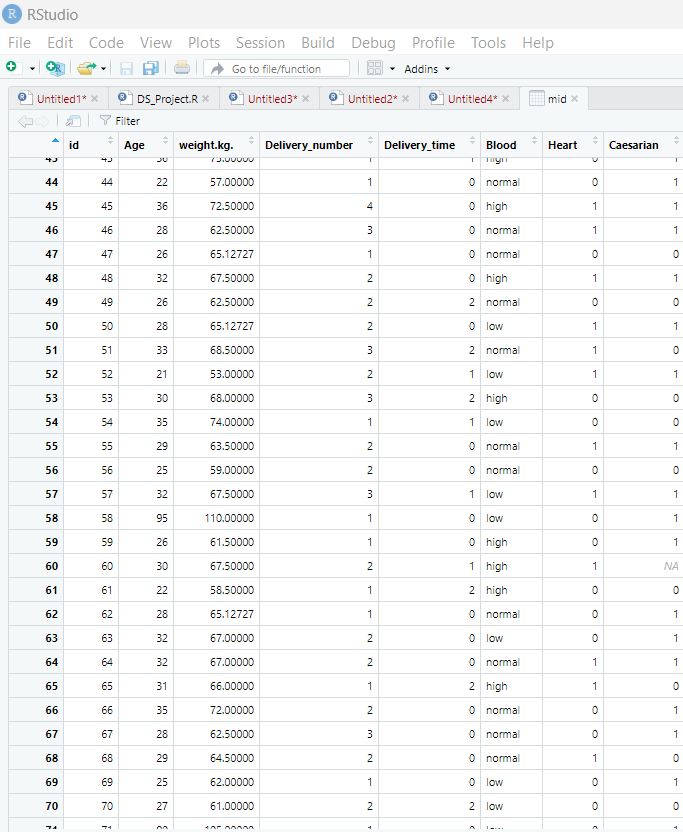
meanweight.kg.

**Replacing Missing Value for weight.kg.**

**Code:**

mid [is.na(mid$weight.kg.),"weight.kg."] = meanweight.kg.

mid

****

**Finding & Replacing Missing Value for weight.kg. (By Median)**

**Code:**

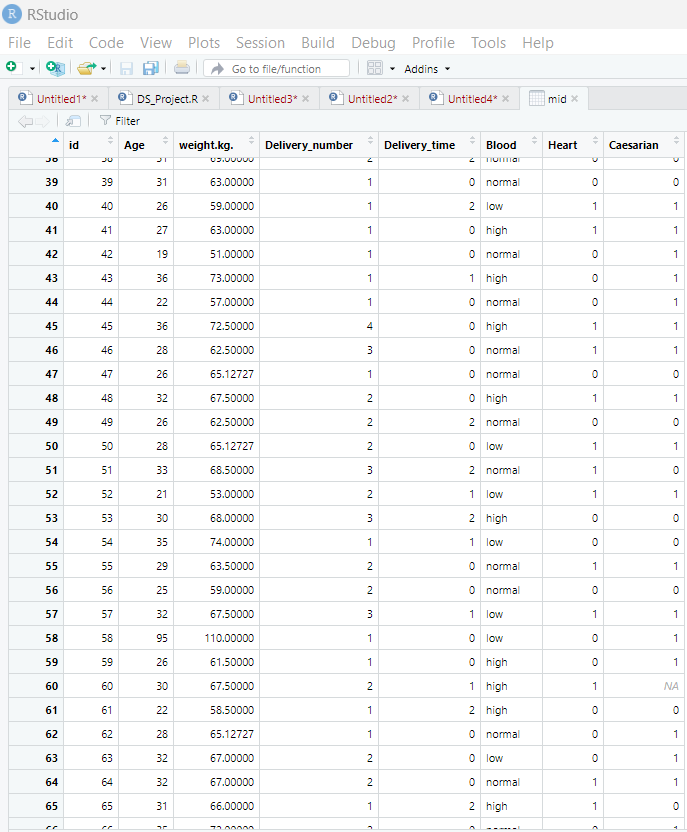
#median weight

medianweight.kg.=median(mid$weight.kg.,na.rm=TRUE)

medianweight.kg.

mid [is.na(mid$weight.kg.),"weight.kg."] = medianweight.kg.

mid



**Finding & Replacing Missing Value for Delivery Number (By Mean)**

**Code:**

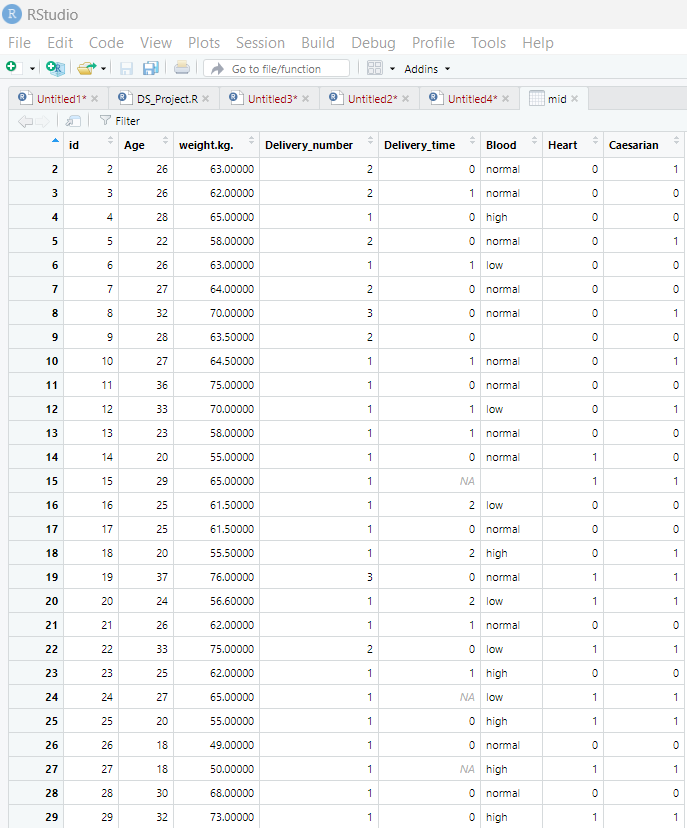
#Missing value For Delivery\_number

meanDelivery\_number=mean(mid$Delivery\_number, na.rm = TRUE)

meanDelivery\_number

mid [is.na(mid$Delivery\_number), "Delivery\_number"]= meanDelivery\_number

mid



**Finding & Replacing Missing Value for Delivery\_number (By Median)**

**Code:**

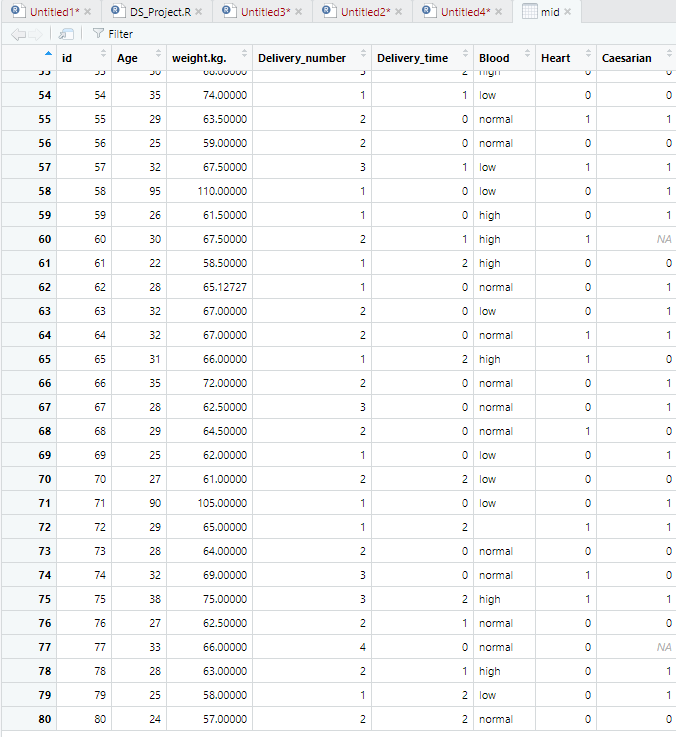
#Median Delivery\_number

medianDelivery\_number=median(mid$Delivery\_number, na.rm = TRUE)

medianDelivery\_number

mid [is.na(mid$Delivery\_number), "Delivery\_number"]= medianDelivery\_number

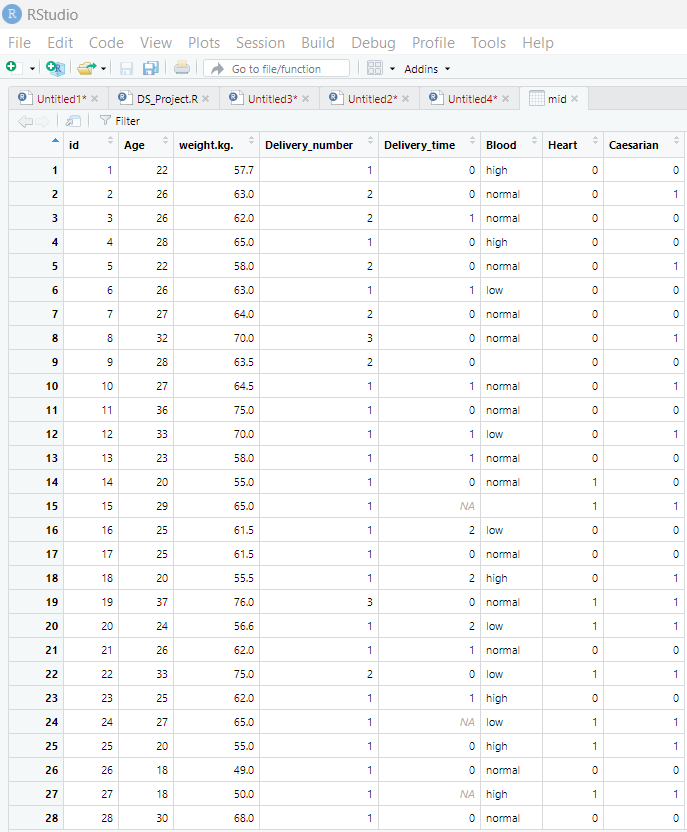
mid



**Finding & Replacing Missing Value for Delivery\_number (By Mode)**

**Code:**

mid$Delivery\_number[is.na(mid$Delivery\_number)] <- Mode(mid$Delivery\_number, na.rm = TRUE)

****

**Finding & Replacing Missing Value for Delivery\_time (By Mean)**

**Code:**

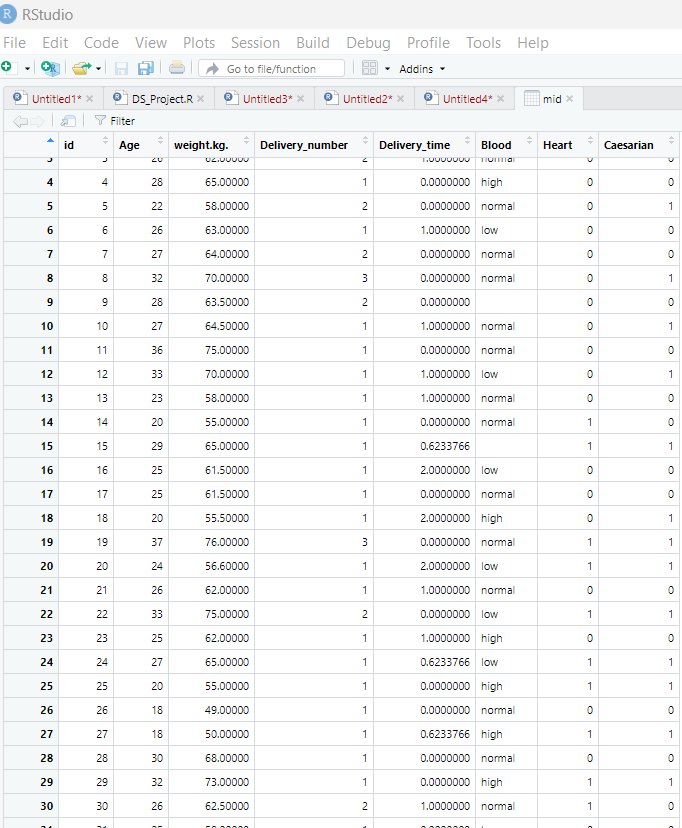
#Missing value For Delivery\_time

meanDelivery\_time=mean(mid$Delivery\_time, na.rm=TRUE)

meanDelivery\_time

mid[is.na(mid$Delivery\_time),"Delivery\_time"] = meanDelivery\_time

mid



**Finding & Replacing Missing Value for Delivery\_time (By Median)**

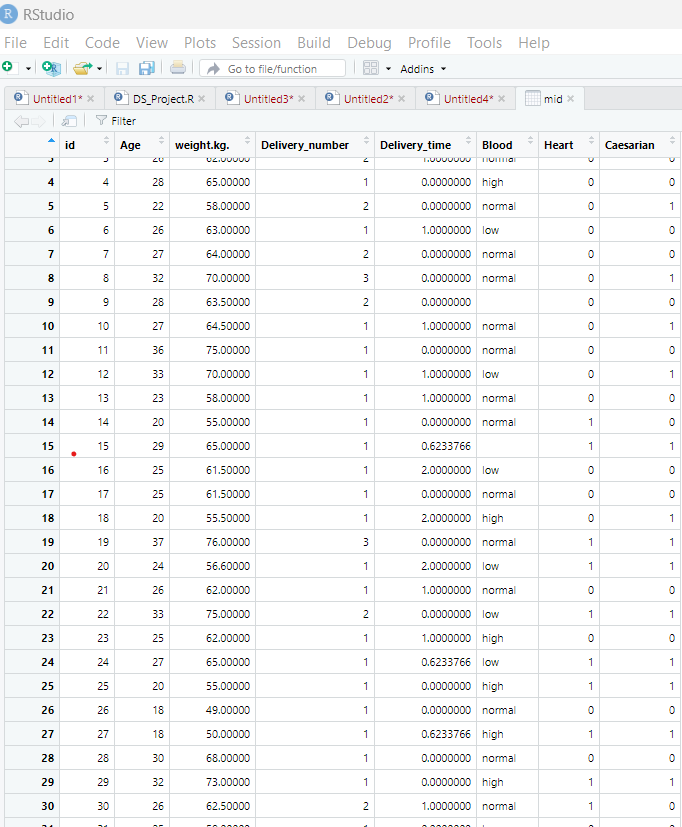
**Code:**

medianDelivery\_time=median(mid$Delivery\_time, na.rm=TRUE)

medianDelivery\_time

mid[is.na(mid$Delivery\_time),"Delivery\_time"] = medianDelivery\_time

mid

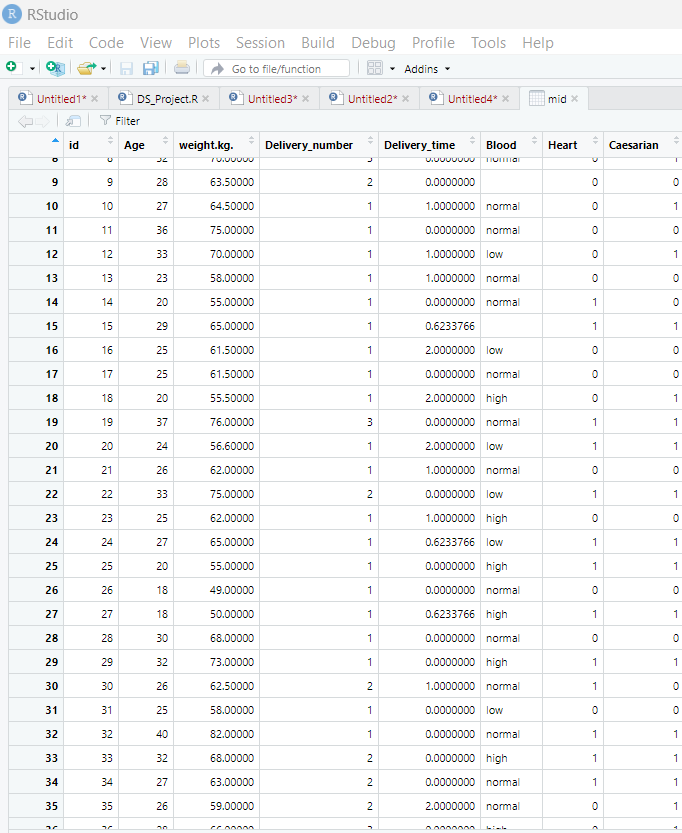


**Finding & Replacing Missing Value for Delivery\_time (By Mode)**

**Code:**

#Mode Delivery time

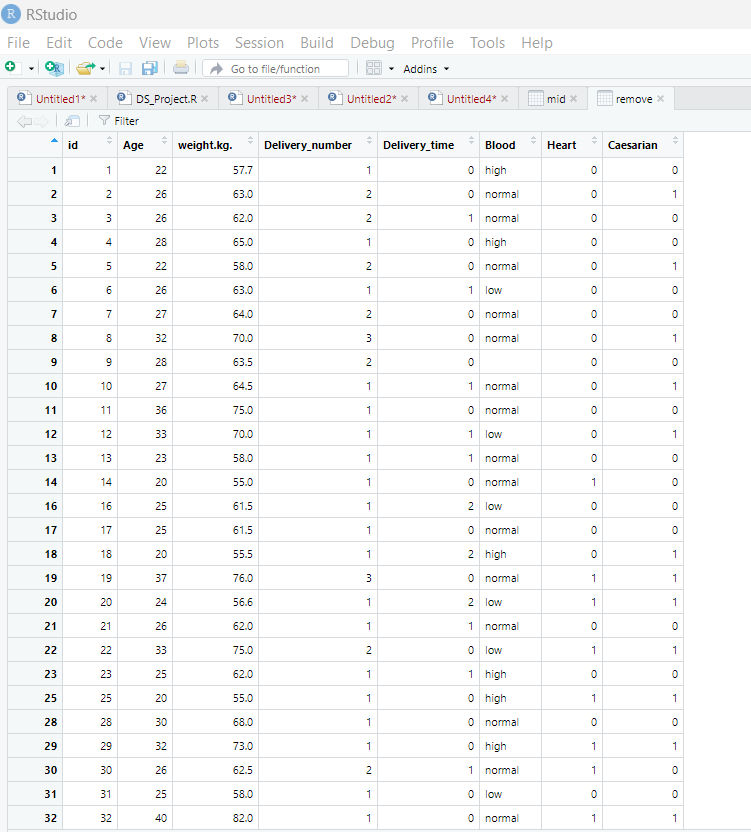
mid$Delivery\_time[is.na(mid$Delivery\_time)] <- Mode(mid$Delivery\_time, na.rm = TRUE)



**Removing Row for Data Preparation**

#Removing Row For data Preparation

remove <- na.omit(mid)



**Finding & Replacing Missing Value for Caesarian (By Mean)**

**Code:**

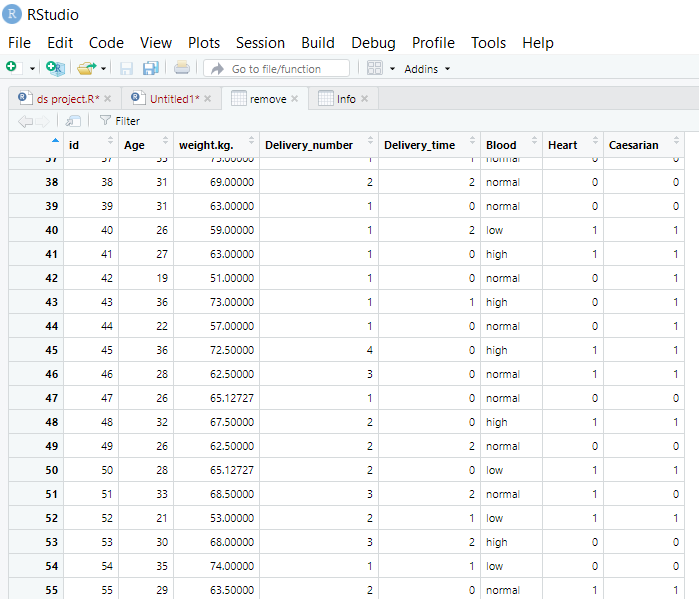
#Missing value For Caesarian

meanCaesarian=mean(mid$Caesarian, na.rm=TRUE)

meanCaesarian

mid [is.na(mid$Caesarian),"Caserian"] = meanCaesarian

mid



**Discussion & Conclusion:**

We will gradually improve the data and process it using R language constructs and methods. The dataset was nicer and cleaner after all data preprocessing techniques were successfully used. However, I wasn't required to use the entire technique for this project. I gained knowledge of the industry's most recent data and data preprocessing. So we need to be careful while doing the data preprocessing a dataset and adding new data to the dataset.