



**Project Tittle:** Caesarian section classification dataset preparation and univariate exploration.

**Date of Submission:** 13/05/23

**Course Title:** Introduction to Data Science

**Section:** [D] **Semester:** Spring 2022-2023

**Course Teacher:** Tohedul Islam

**Submitted By:**

**NAME:** Tanjim, Samiul Arif

**ID:** 20-42694-1

**Dataset Description:** The dataset contains information about caesarian section results of 80 pregnant women with the most important characteristics of Delivery problems in the medical field.

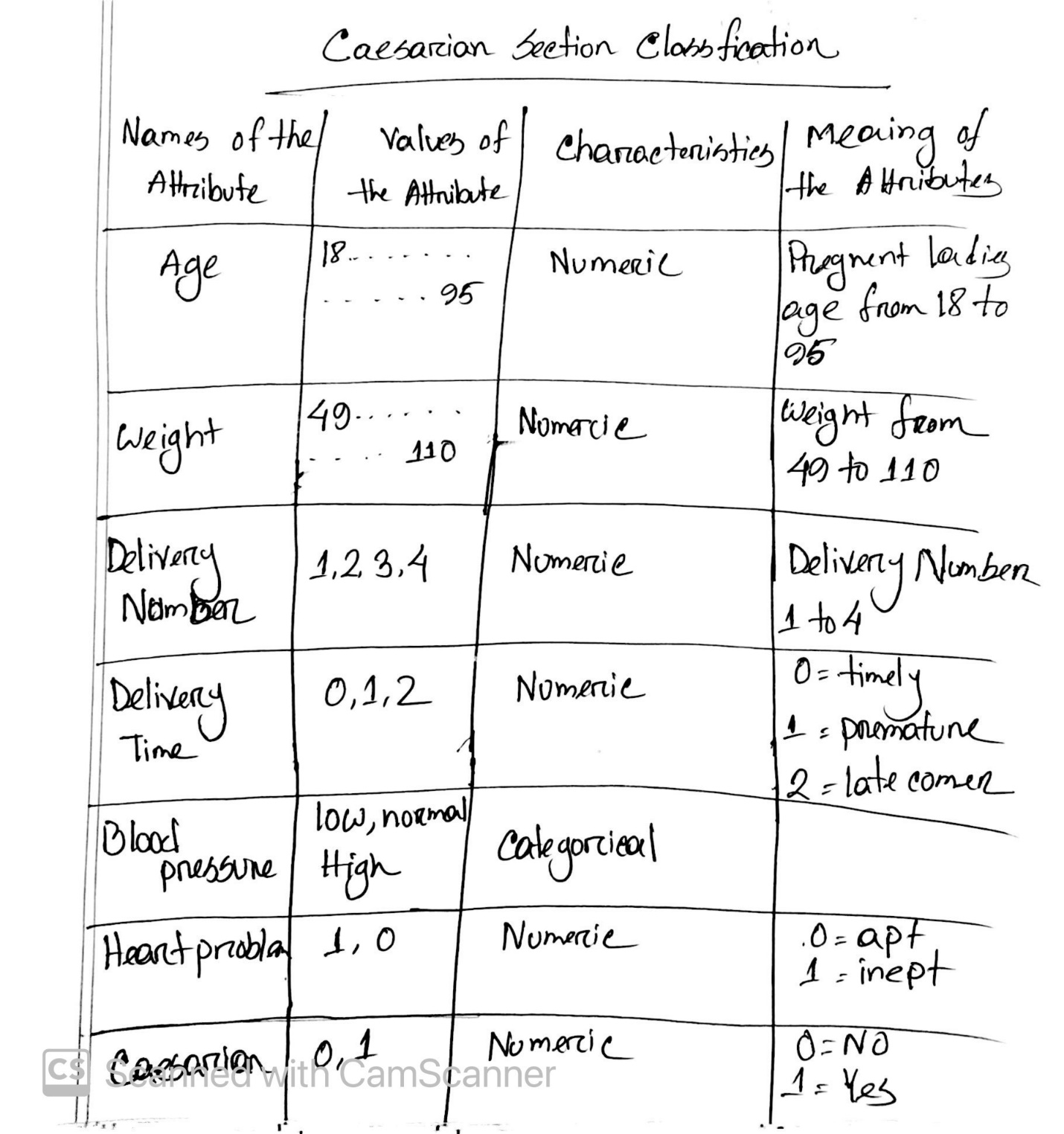


Fig 1: Attribute description of the dataset

**Data Preparation**

**Data Types and Conversion:** In this dataset, Blood pressure of pregnant ladies are described in categorical value. Different data type in a dataset imposes different restriction on the evaluation of dataset.

**Transformation:** The blood pressure attribute is expected to be numeric. For comparing the values of different attributes and calculate numerical value it is needed to transform categorical value to numerical values.

**Annotating dataset:** Typically, annotation includes adding descriptive labels to variable names and value labels to the codes used for categorical variable. In this dataset, we have labeled blood pressure into numerical representation using ‘low as 0’, ‘normal as 1’ and ‘high as 2’. The factor () function is used to create value labels for categorical variables.

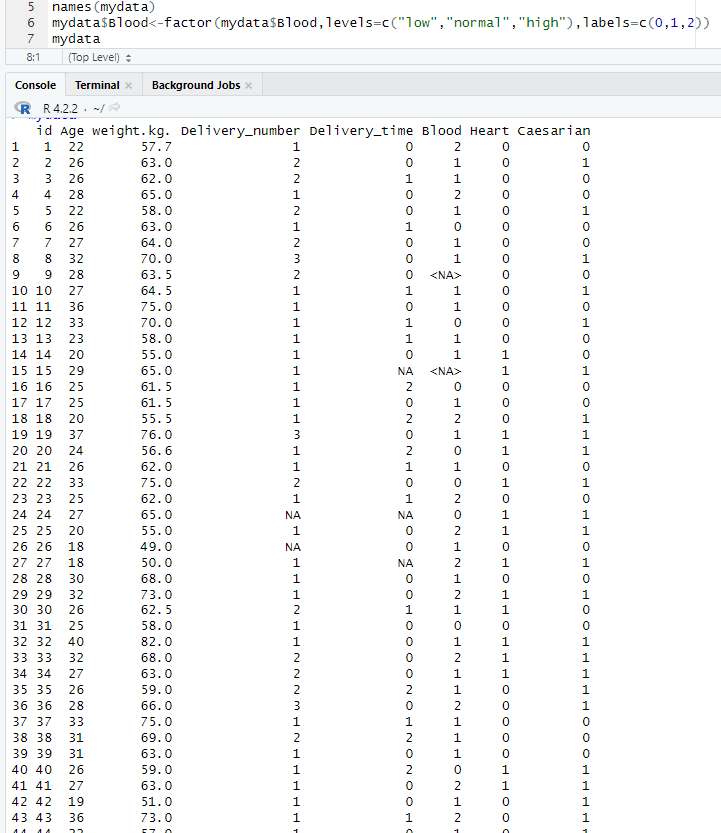


Fig 2: Labeling blood pressure using numeric value

**Outlier:** In this dataset, we have found that there are two instances where the age of the pregnant ladies is around 90 and 95. For this dataset, there data points are significantly different from the rest of dataset. Here we have removed those instances having outlier.

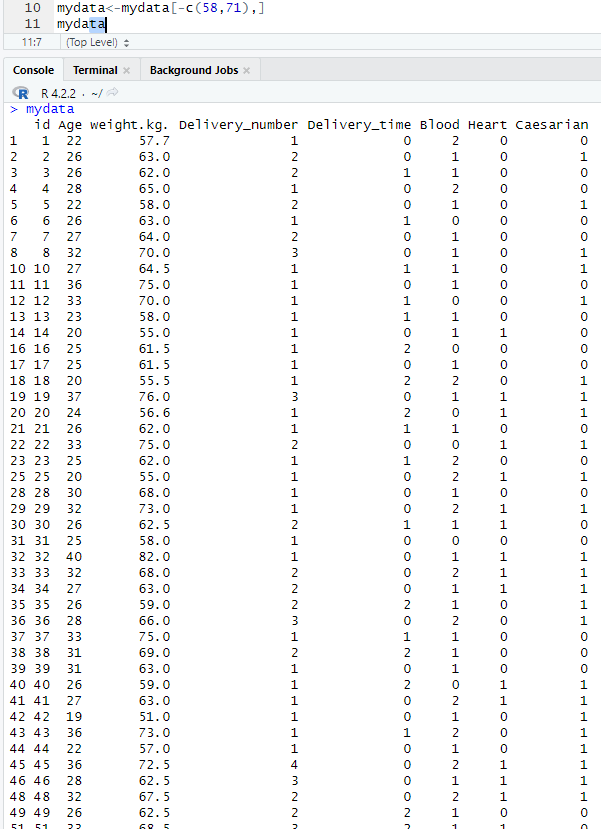


Fig 3: Removing Outlier from the dataset

**Finding missing values:** For counting numbers of null values in each column, we have used **colSums(is.na(mydata)** function.

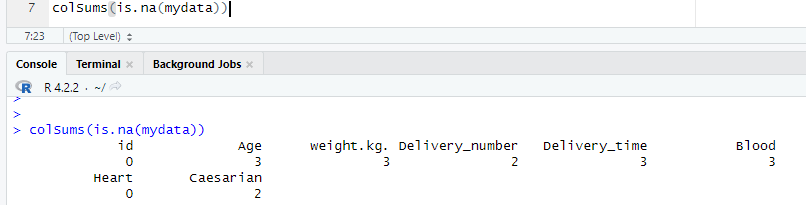


Fig 4: Counting missing values

**Removing missing values:** We can deal with missing values in several way. For preparing this dataset we have removed all the instance having null values. To remove null values from the dataset **na.omit(mydata)** function was used.

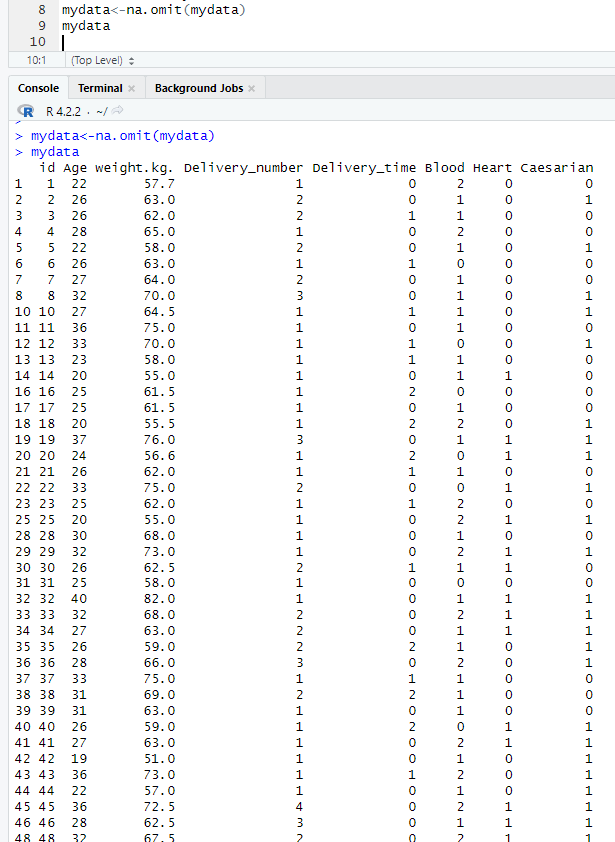


Fig 5: Removing null values

**Structure of the new dataset:** To see the modified dataset summery of the structure **str(mydata)** function was used.

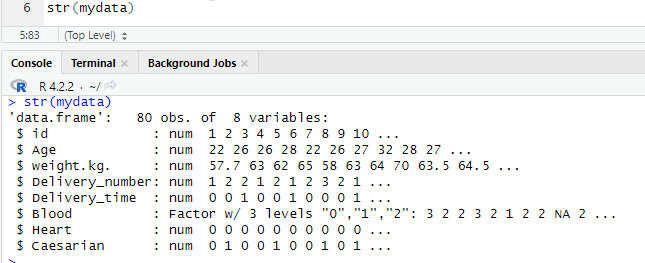


Fig:6 Structure summary of new dataset

**Sampling dataset:** To take sample of the dataset we have used **sample\_n(mydata,3)** function. Here 3 indicated how many rows of sample it will generate.

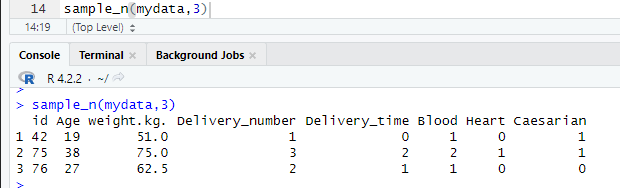
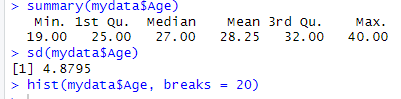


Fig 7: Sample dataset

**Univariate Exploration**

**Univariate Exploration of Age attribute:** We have used **summary(mydata$age)** function to find the descriptive statistics and **sd (mydata$age)** function to find the standard variance of age attribute and also **hist (mydata$age, breaks =20)** function to plot the age attribute where breaks=20 means that the plotting will show in 20 column.





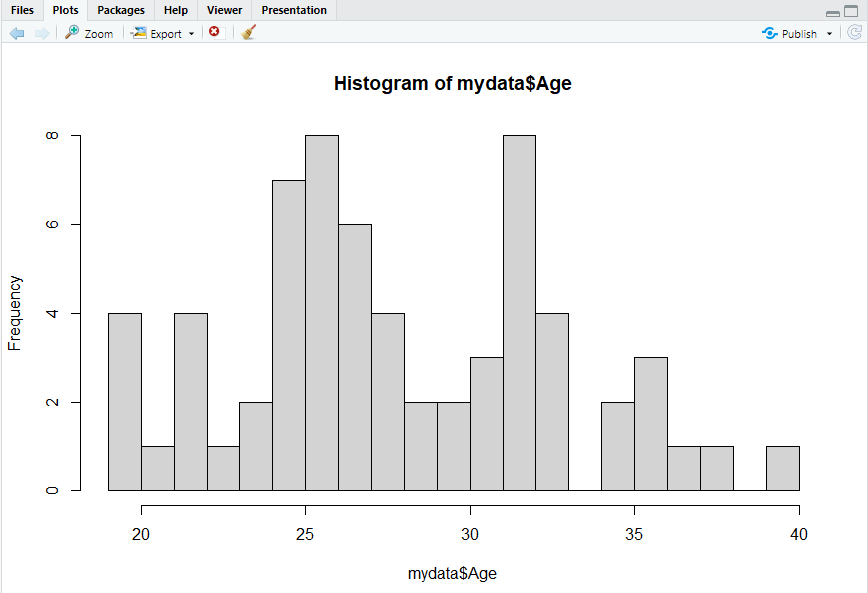
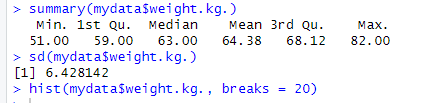


Fig 8: Descriptive statistics, standard variance and histogram of **Age** attribute

**Here, we can see that, there is no data available for age around 33-34 and 38-39**

**Univariate Exploration of weight.kg attribute:** We have used **summary(mydata$weight.kg.)** function to find the descriptive statistics and **sd (mydata$weight.kg.)** function to find the standard variance of weight attribute and also **hist (mydata$weight.kg., breaks =20)** function to plot the weight attribute where breaks=20 means that the plotting will show in 20 column.





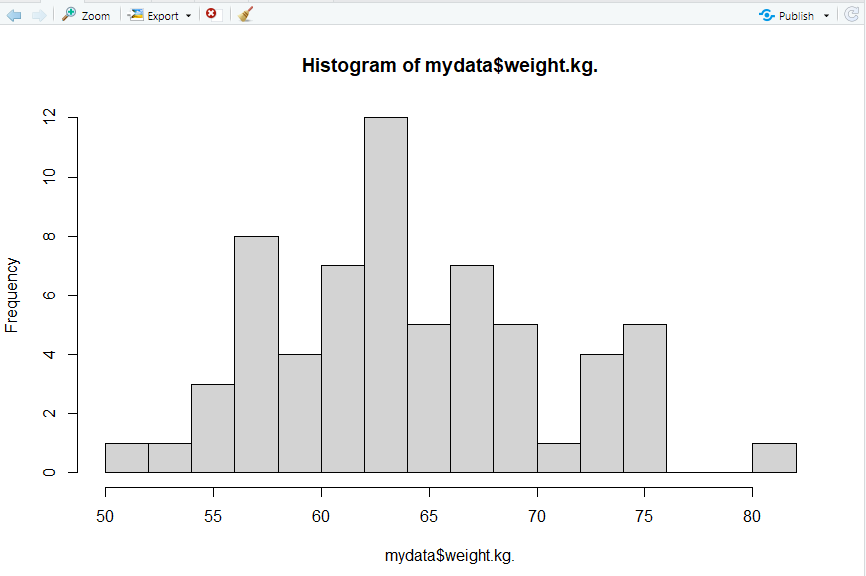
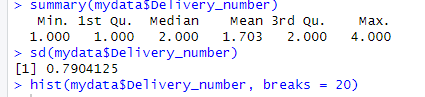


Fig 9: Descriptive statistics, standard variance and histogram of **weight.kg** attribute

**Here we can see that, there is no data available for weight.kg from 76-80 kg.**

**Univariate Exploration of Delivery Number attribute:** We have used **summary(mydata$Delivery\_number)** function to find the descriptive statistics and **sd (mydata$Delivery\_number)** function to find the standard variance of delivery number attribute and also **hist (mydata$Delivery\_number, breaks =20)** function to plot the age attribute where breaks=20 means that the plotting will show in 20 column.





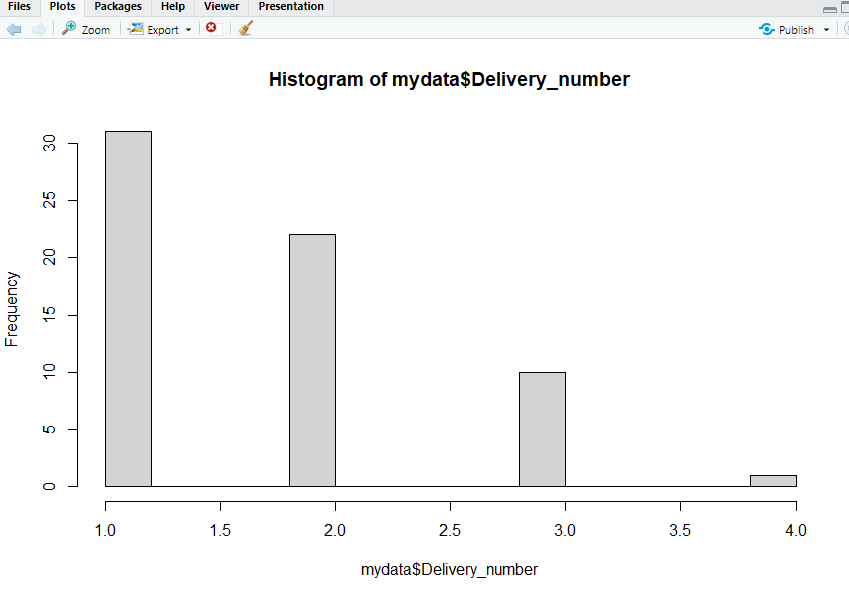
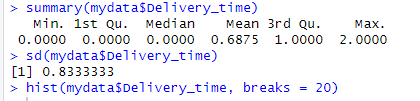


Fig 10: Descriptive statistics, standard variance and histogram of **Delivery\_number** attribute

**Univariate Exploration of Delivery\_time attribute:** We have used **summary(mydata$Delivery\_time)** function to find the descriptive statistics and **sd (mydata$Delivery\_time)** function to find the standard variance of Delivery\_time attribute and also **hist (mydata$Delivery\_time, breaks =20)** function to plot the age attribute where breaks=20 means that the plotting will show in 20 column.





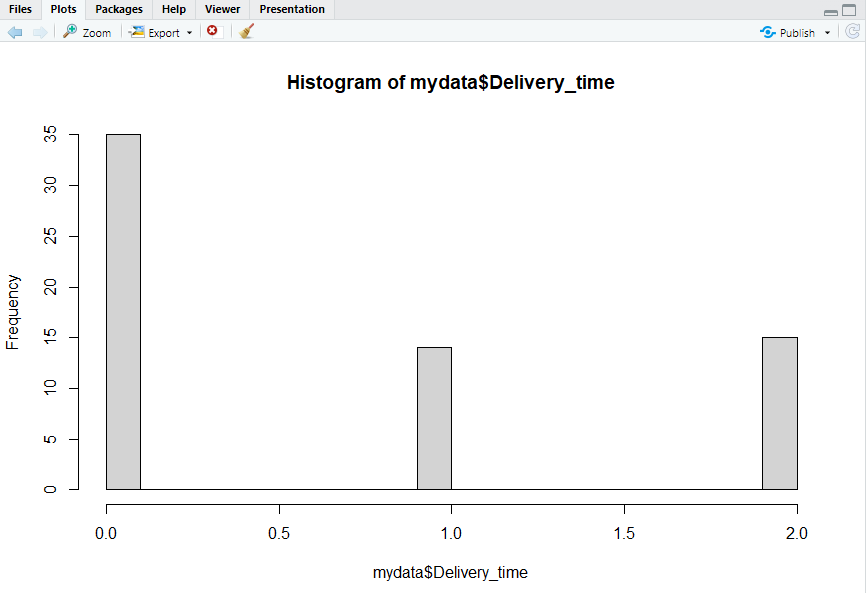
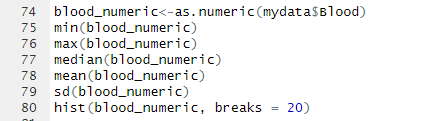
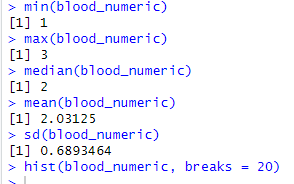


Fig 11: Descriptive statistics, standard variance and histogram of **Delivery\_time** attribute

**Univariate Exploration of Blood attribute:** As we have factorize the blood pressure attribute from categorical to numerical values, we have to use **as.numeric(mydata$Blood)** function to summarize the descriptive statistics. We have used **min(blood\_numeric), max(blood\_numeric), median(blood\_numeric), mean(blood\_numeric)** function to find the descriptive statistics and **sd(blood\_numeric)** function to find the standard variance of Blood pressure attribute and also **hist(blood\_numeric, breaks =20)** function to plot the age attribute where breaks=20 means that the plotting will show in 20 column.





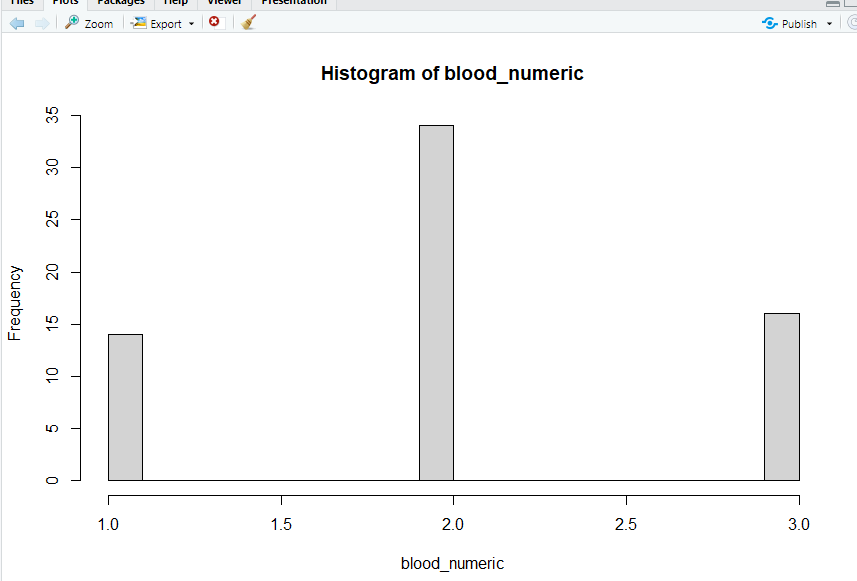
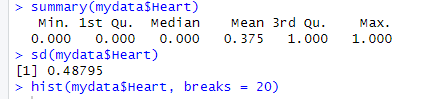


Fig 12: Descriptive statistics, standard variance and histogram of **Blood** Attribute

**Univariate Exploration of Heart attribute:** We have used **summary(mydata$Heart)** function to find the descriptive statistics and **sd (mydata$Heart)** function to find the standard variance of Heart problem attribute and also **hist (mydata$Heart, breaks =20)** function to plot the age attribute where breaks=20 means that the plotting will show in 20 column.





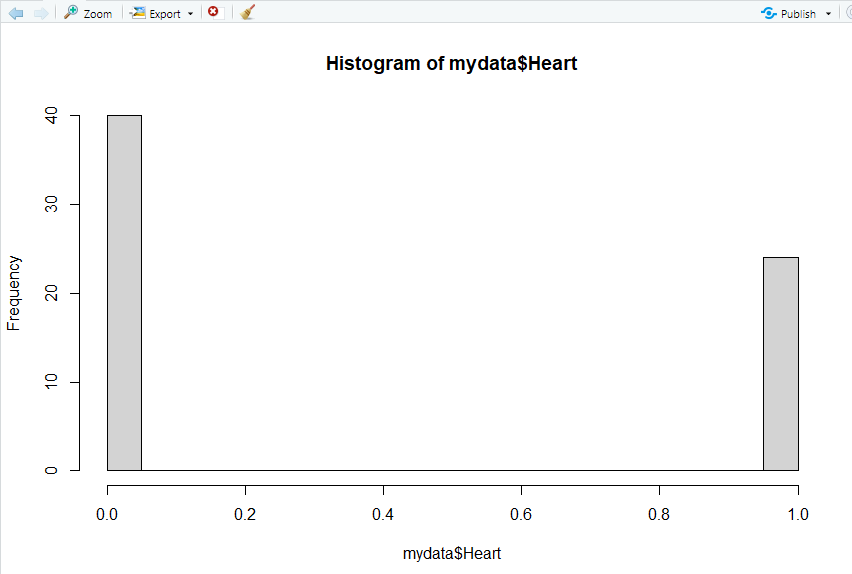
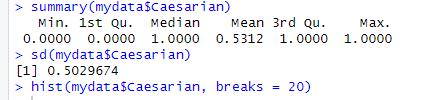


Fig 13: Descriptive statistics, standard variance and histogram of **Heart** Attribute

**Univariate Exploration of Caesarian attribute:** We have used **summary(mydata$Caesarian)** function to find the descriptive statistics and **sd (mydata$Caesarian)** function to find the standard variance of Caesarian attribute and also **hist (mydata$Caesarian, breaks =20)** function to plot the age attribute where breaks=20 means that the plotting will show in 20 column.





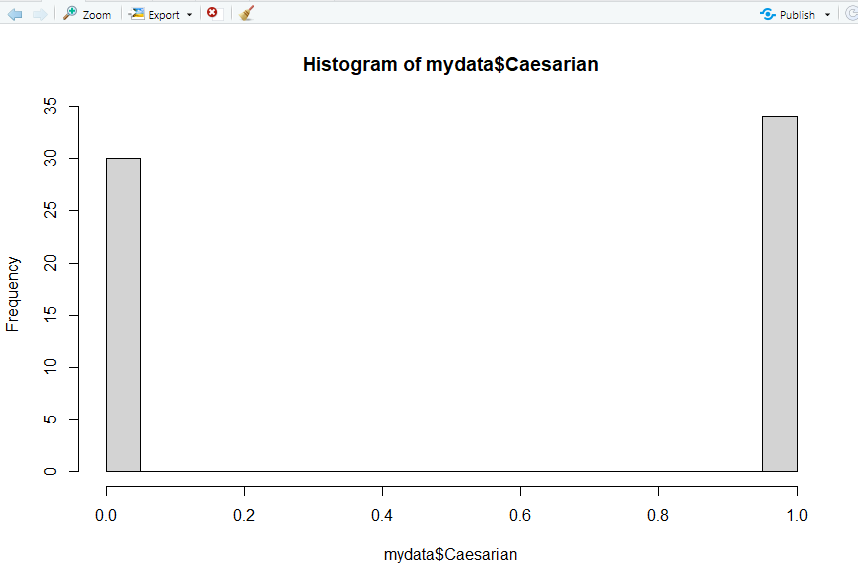


Fig 14: Descriptive statistics, standard variance and histogram of **Caesarian** Attribute

**Outcomes**

* Almost 50% of all pregnant women have normal blood pressure. The proportions of women with low and high blood pressure are roughly the same.
* Around 40% of women had only had one previous delivery.
* More than 65% of pregnant women deliver on time. The proportion of women who have premature or late deliveries is nearly the same.
* The majority of women has no heart problem
* About 50% of women had a caesarian section (53.12%)