**Mid Term Project**

**Introduction to Data Science**

**Section: A**

**Submitted by:**

|  |  |
| --- | --- |
| **Name** | **ID** |
| **RABBI HOSSEN** | **20-44220-3** |
| **UDDIN, MD BORHAN** | **20-44002-2** |

**Submitted to: TOHEDUL ISLAM**

**Dataset Description:**

"This 1,319-sample dataset explores various reasons behind heart attacks, providing information about patients who have experienced this condition."  
we utilize a dataset comprising 1,319 samples, each with nine distinct fields. Sixth of these fields serve as input parameters, including age, gender, heart rate, systolic and diastolic blood pressure and blood sugar (glucose) levels. The seventh field is the output parameter/target attribute, classifying the presence of a heart attack into two categories: "negative" for the absence of a heart attack and "positive" for the presence of a heart attack.  
Analyzing this dataset may lead to critical insights and advancements in the field of cardiovascular health, ultimately helping to reduce the global burden of CVD-related deaths and improve public health outcomes.

**Convert XLSX to CSV and Import CSV and print the dataset:**

**A screenshot of a computer

Description automatically generated**

**Description:** The code is for reading the dataset and to prove the successful conversion of a xlsx file to csv file. Also, the output of the dataset by using the print()

**Column Name:**

names(data)

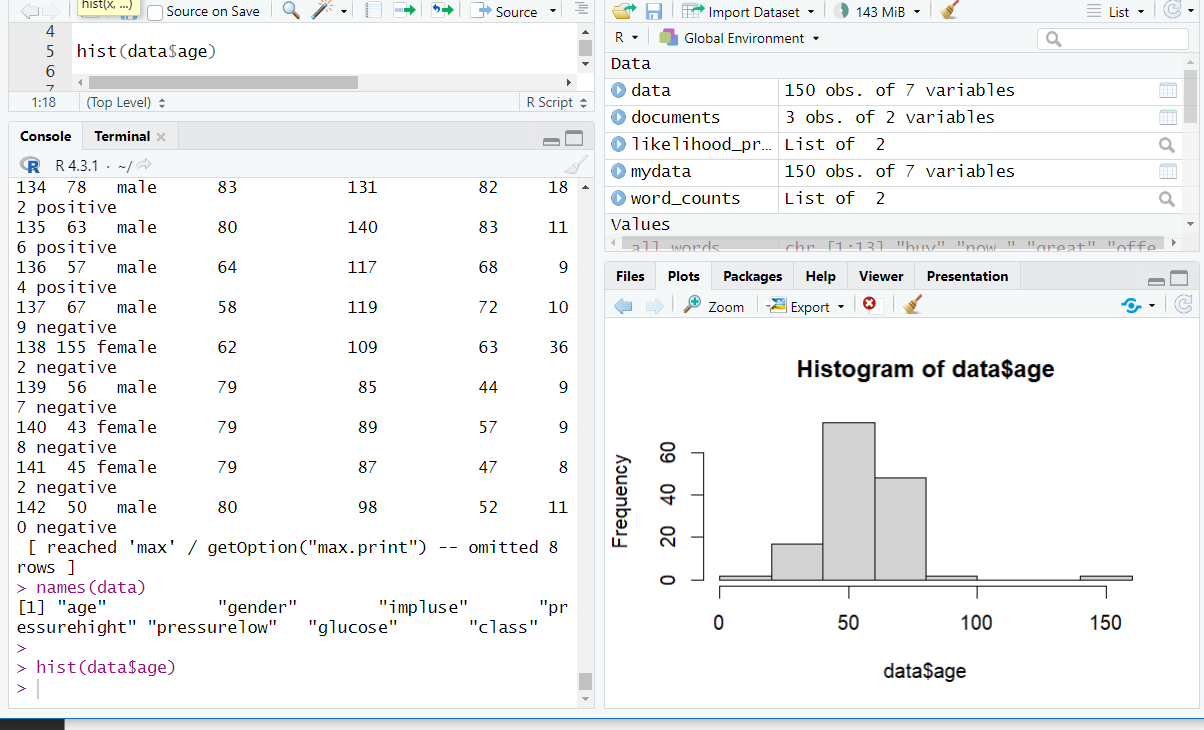
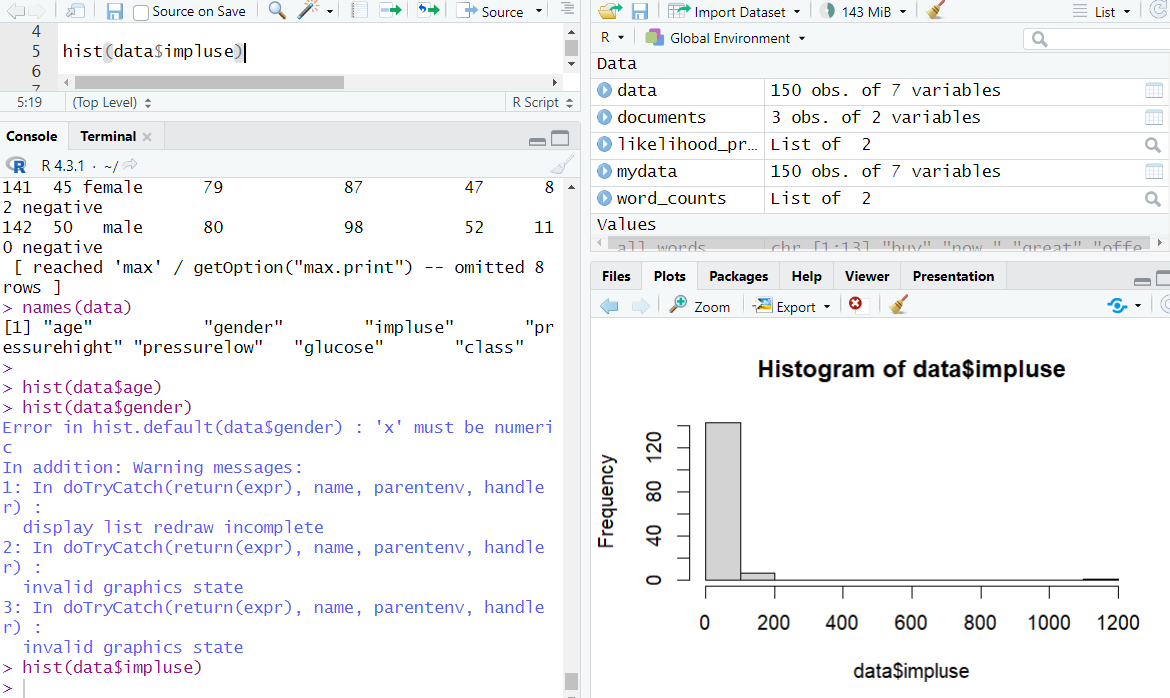
**A screenshot of a computer

Description automatically generated**

**Description:**

Use ‘Names()’ function to see columns name of the dataset.

**Exploration:**

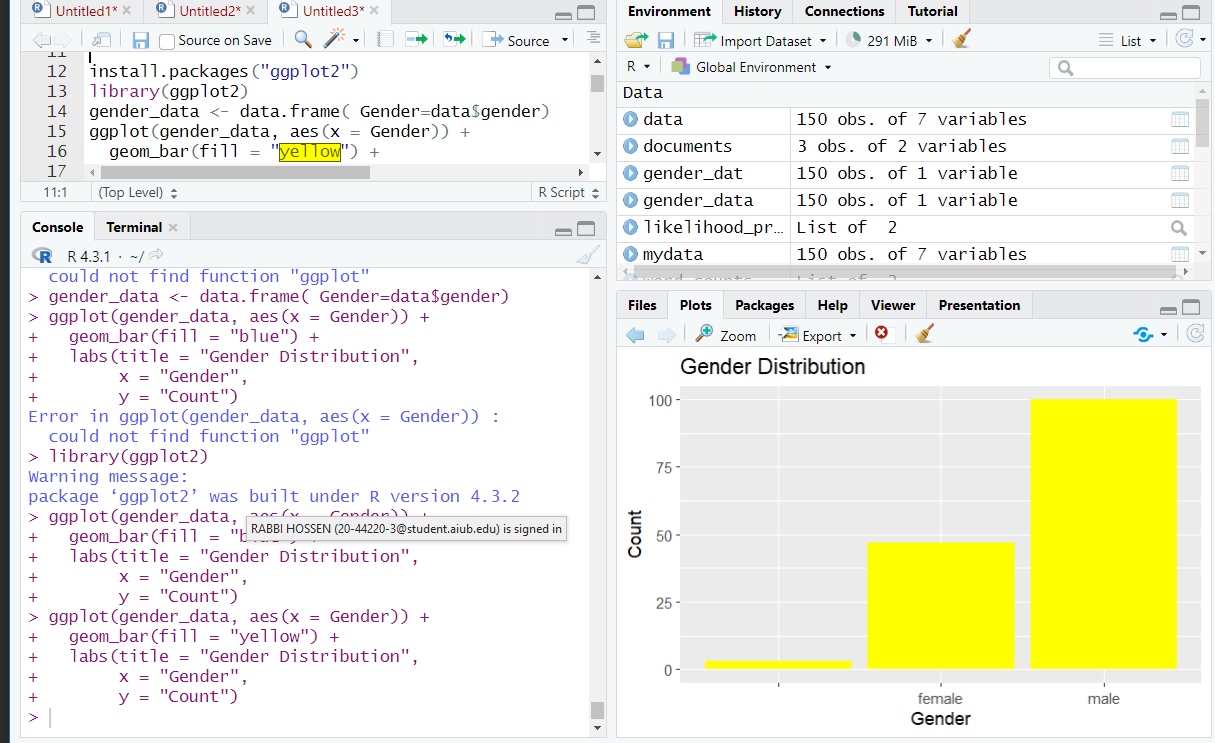
 ****

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated 

A screenshot of a computer

Description automatically generated

**Description:**

Use hist() function to visualize attributes of numerical values and use barplot to visualize categorical values (gender and class). We use “ggplot2” library to demonstrate barchart.

**Data-Quality (Check Data-Completeness):**

A screenshot of a computer code

Description automatically generated

**Description:**

is.na indicates which elements are missing.

To check if the dataset is complete or not we use ‘colMeans()’ function. If not complete than it shows less than 1 value but if it is complete than it shows the value 1. We find that there is no complete data in our dataset

**Conversion:**

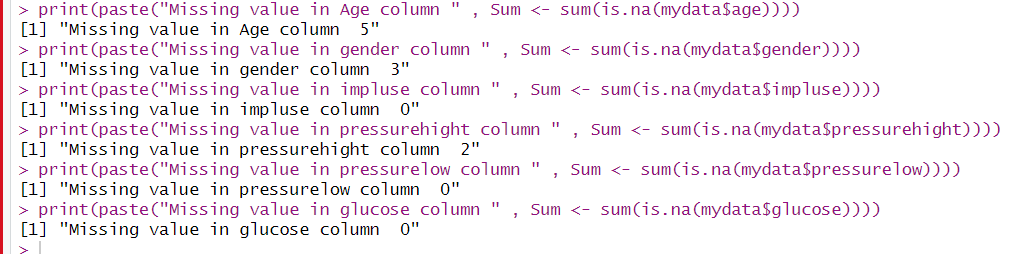
**A screenshot of a computer

Description automatically generated**

**Description:**

We convert categorical value to numerical value for getting accurate mathematic calculation. Levels and label are used for transform characters to integer and bind them and “factor()” function  is used to encode a vector as a factor.

**Missing value finding:**

****

**Description:**

using is.na() function to check is there any missing value or not.

**Missing value handling (Remove null value):**

A screenshot of a computer

Description automatically generated

**Description:**

We remove all null values row in the dataset using “na.omit()” function.

**Missing value handling (Replace by mean value):**

**A screenshot of a computer program

Description automatically generated**

**Description:**

we use “mean()” function to replace the null/missing value for age and pressurehigh column.

**Missing value handling (Replace by median value):**

A screenshot of a computer

Description automatically generated

**Description:**

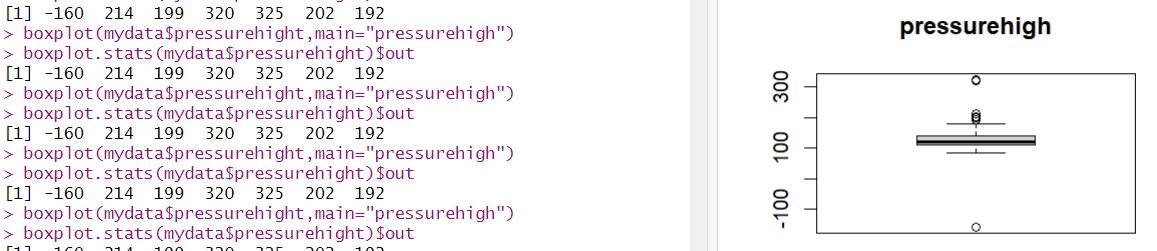
We Use “median()” function to replace the null/missing value for age and pressurehight column.

**Outliers (Detect and Handle):**

A white background with black text

Description automatically generatedA computer screen with a white background

Description automatically generated

 A screenshot of a computer

Description automatically generated