A blue circle with text and images

Description automatically generated**AMERICAN INTERNATIONAL**

**UNIVERSITY-BANGLADESH**

408/1, Kuratoli, Khilkhet, Dhaka 1229, Bangladesh

|  |  |  |
| --- | --- | --- |
| Project Title: Data Preparation and analysis |  |  |
| Project No: Mid 01 |  | Date of Submission: 24/08/24 |
| Course Title: INTRODUCTION TO DATA SCIENCE |  |  |
| Course Code: 00489 |  | Section: A |
| Semester: Summer |  | Course Teacher: TOHEDUL ISLAM |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **Name** | **ID** | **Program** | **Signature** |
| 1 | **Md. Omar Faruk Sakib** | 21-45077-2 | BSc CSE |  |
| 2 | **Md. Abu Hojifa** | 21-45081-2 | BSc CSE |  |
|  |  |  |  |  |

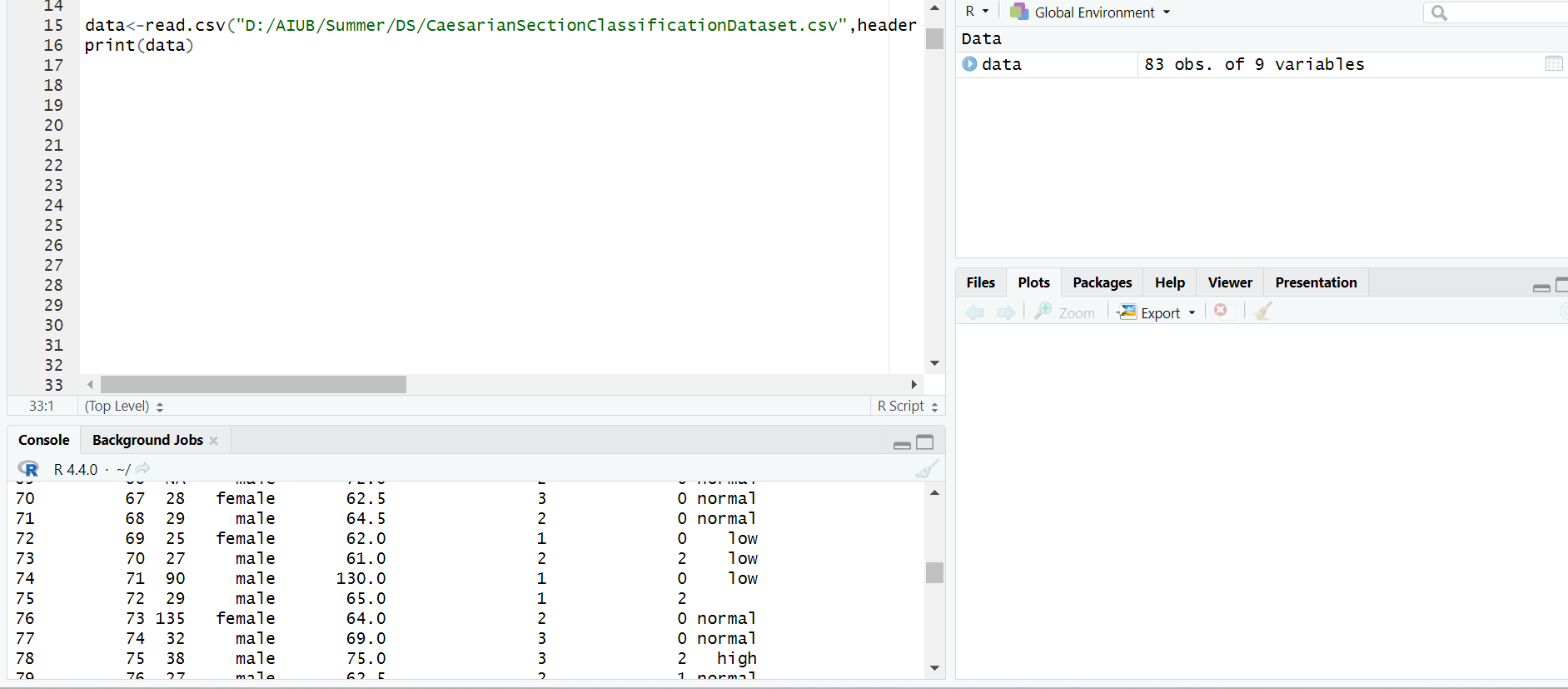
|  |  |  |
| --- | --- | --- |
| ***Faculty use only*** |  | |
|  | **Marks Obtained** |  |
|  |
|  |
|  | **Total Marks** |  |
|  |  |  |

1. **Introduction:**

This Dataset provides a comprehensive view of factors influencing the decision to perform a caesarian section during childbirth. Comprising data on 83 pregnant women, the dataset includes attributes like age, delivery number, delivery time, blood pressure, and the presence of heart problems. These attributes offer insights into the health and medical history of the women, which are crucial considerations for medical professionals when determining the safest method of delivery. The dataset's target attribute, indicating whether a caesarian section was performed, serves as a valuable label for predictive modeling tasks. Analyzing this dataset can potentially lead to the development of predictive models that assist healthcare providers in making informed decisions about childbirth procedures, ultimately contributing to improved maternal and neonatal outcomes.

1. **Import Dataset:**

Here we use the read.csv () function to import the CSV file title “CaesarianSectionClassificationDataset.csv” from the directory "D:\AIUB\Summer\DS\" into a variable named dsp.



**Title-Importing Dataset**

1. **Data Exploration:**

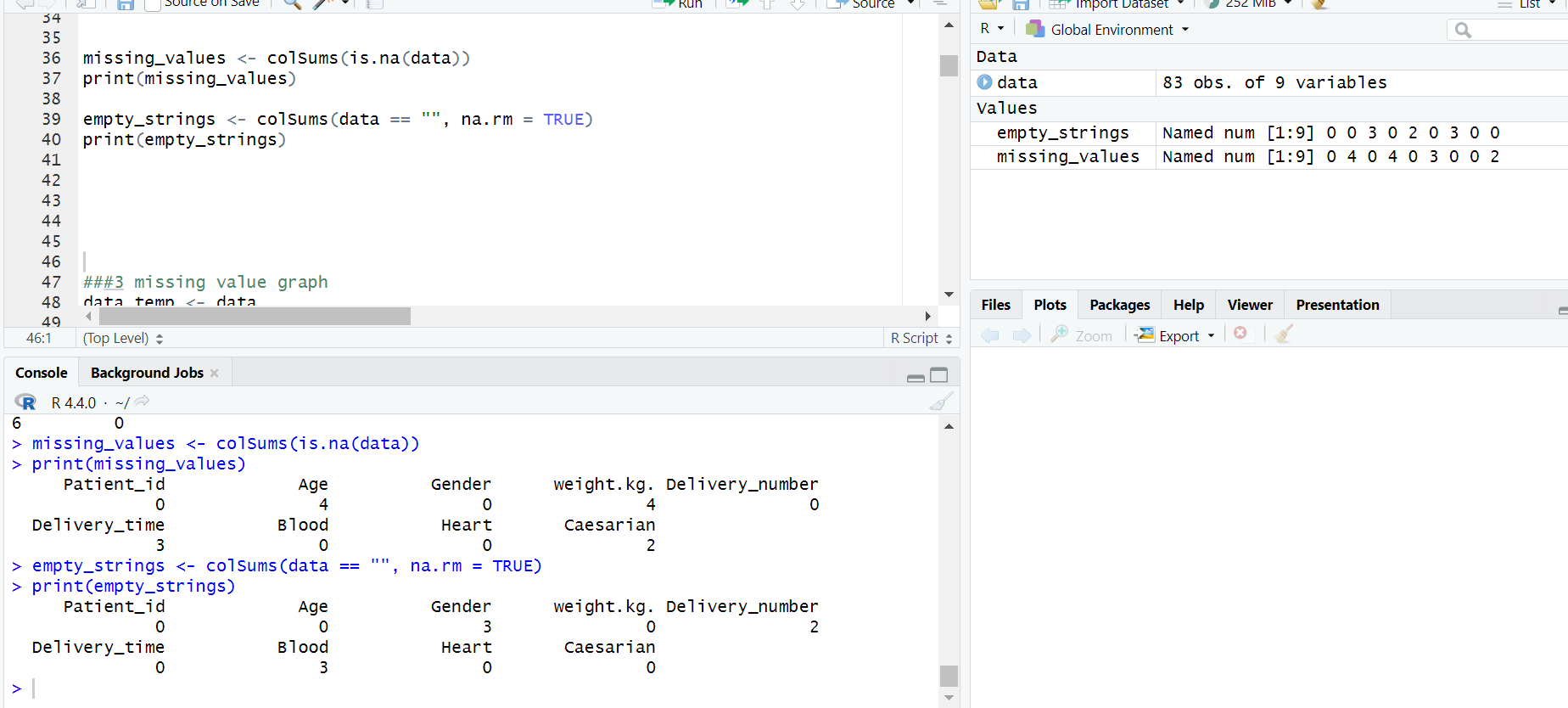
The initial examination of the dataset to understand its structure and content. It involves looking at the dataset's size (number of observations and variables), and key variables such as age, gender, weight, delivery number, blood type, and heart condition. This analysis helps identify patterns, relationships, and any potential issues in the data.A screenshot of a computer

Description automatically generated

**Title:** **Exploratory Data Analysis**

1. **Finding Missing Values:**

This code snippet is for identifying missing values within the dataset ‘data. It includes commands for counting missing values and empty strings. Focuses on identifying missing values within the dataset. Missing data can affect the quality and outcome of data analysis, so it’s crucial to detect where data is missing. use of specific commands to count and locate these missing values.



**Title:** **Missing Values of the data set**

A screenshot of a computer

Description automatically generated

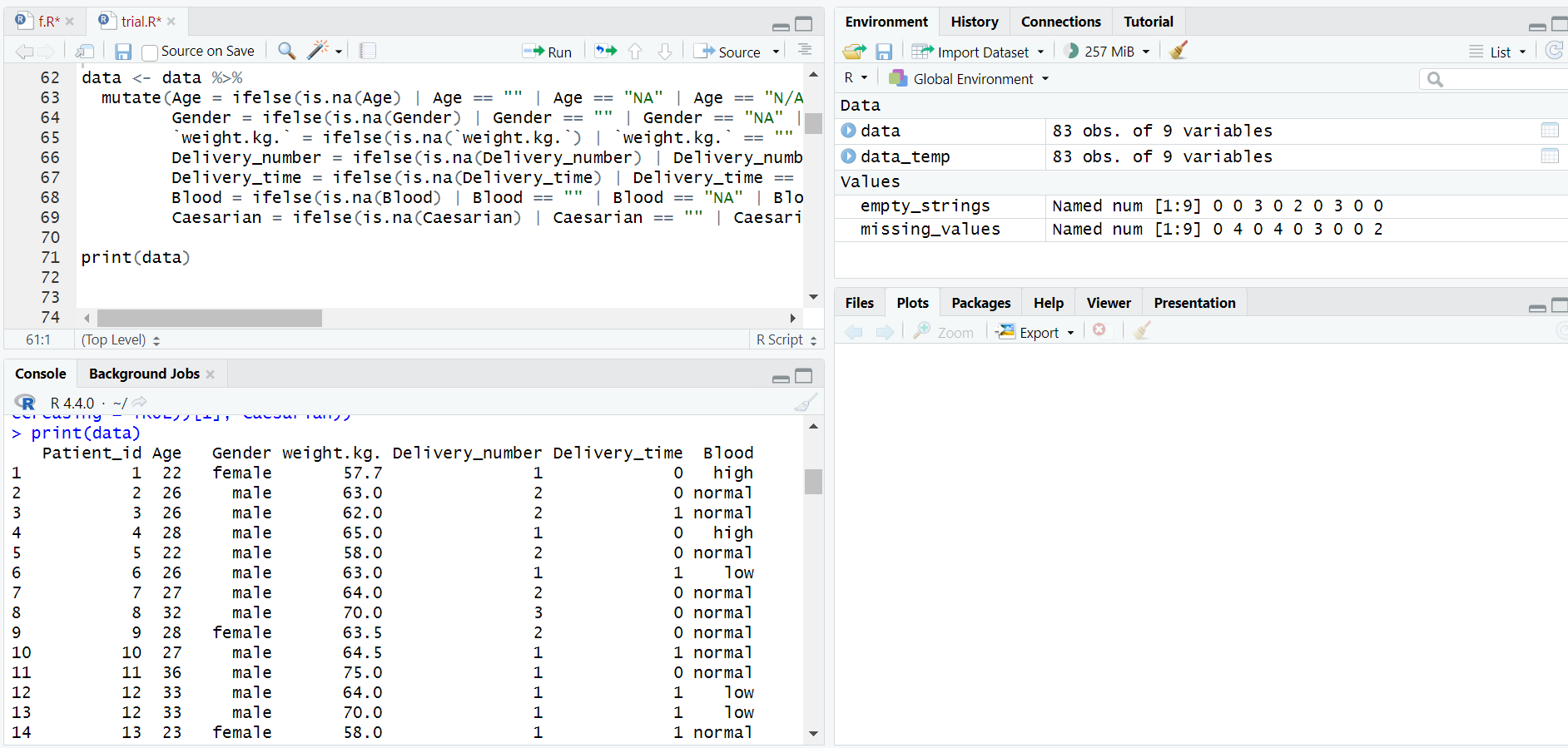
**Title:** **Missing Values graph of the data set**

1. **Applying Methods to Handle Missing Values:**

Various techniques to address missing data, such as discarding instances with missing values, using imputation methods (mean, median, mode, or advanced techniques like machine learning-based imputation), or replacing missing values with placeholders. These methods are essential to maintain data integrity, ensure accurate analysis, and prevent biased results, depending on the nature and extent of the missing data in the dataset.

1. **Discard Instances:**

Handling missing values, empty strings, and placeholders

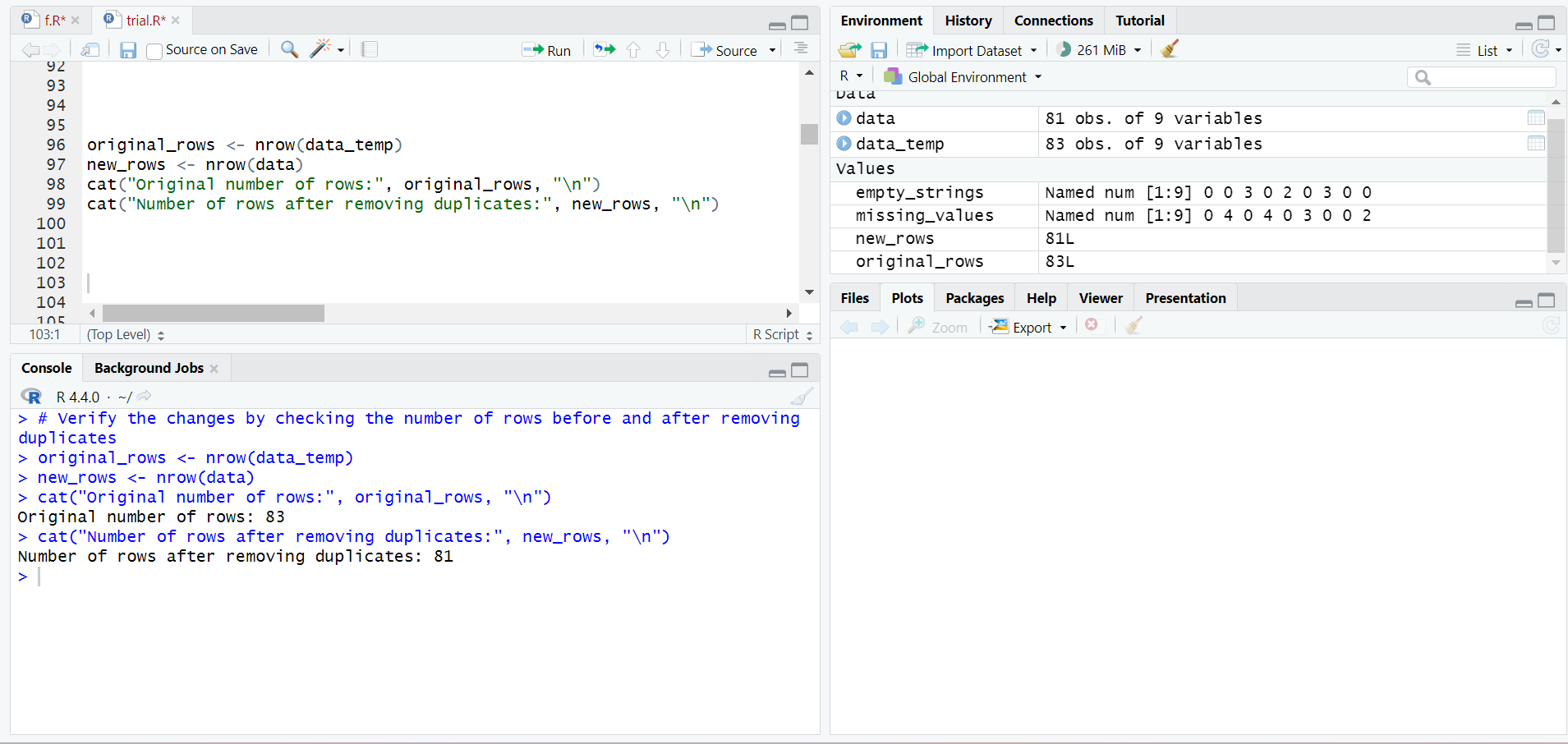


**Title: Removing the** **Missing Values of the data set**

A screenshot of a computer

Description automatically generated

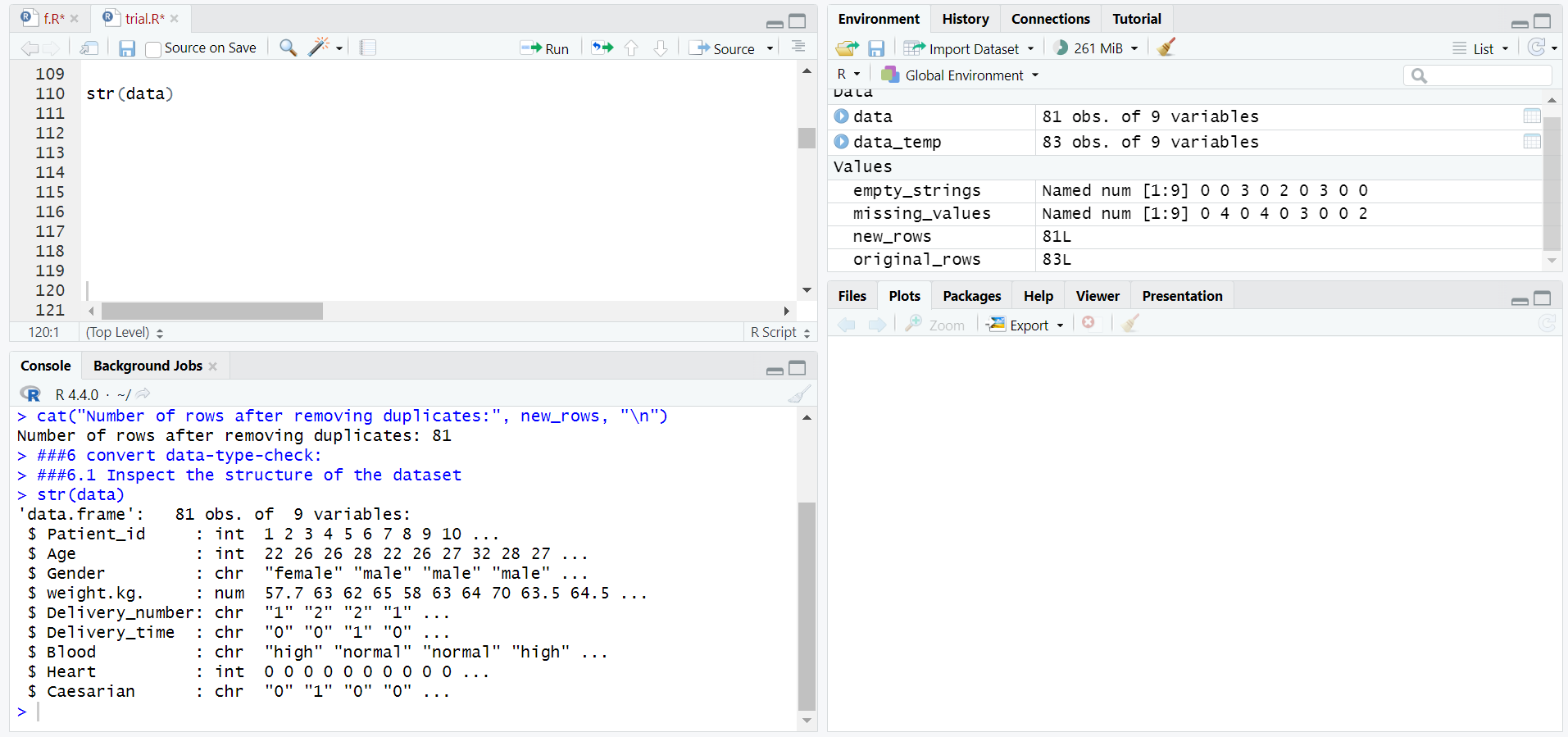
**Title:** **Remove duplicate rows**



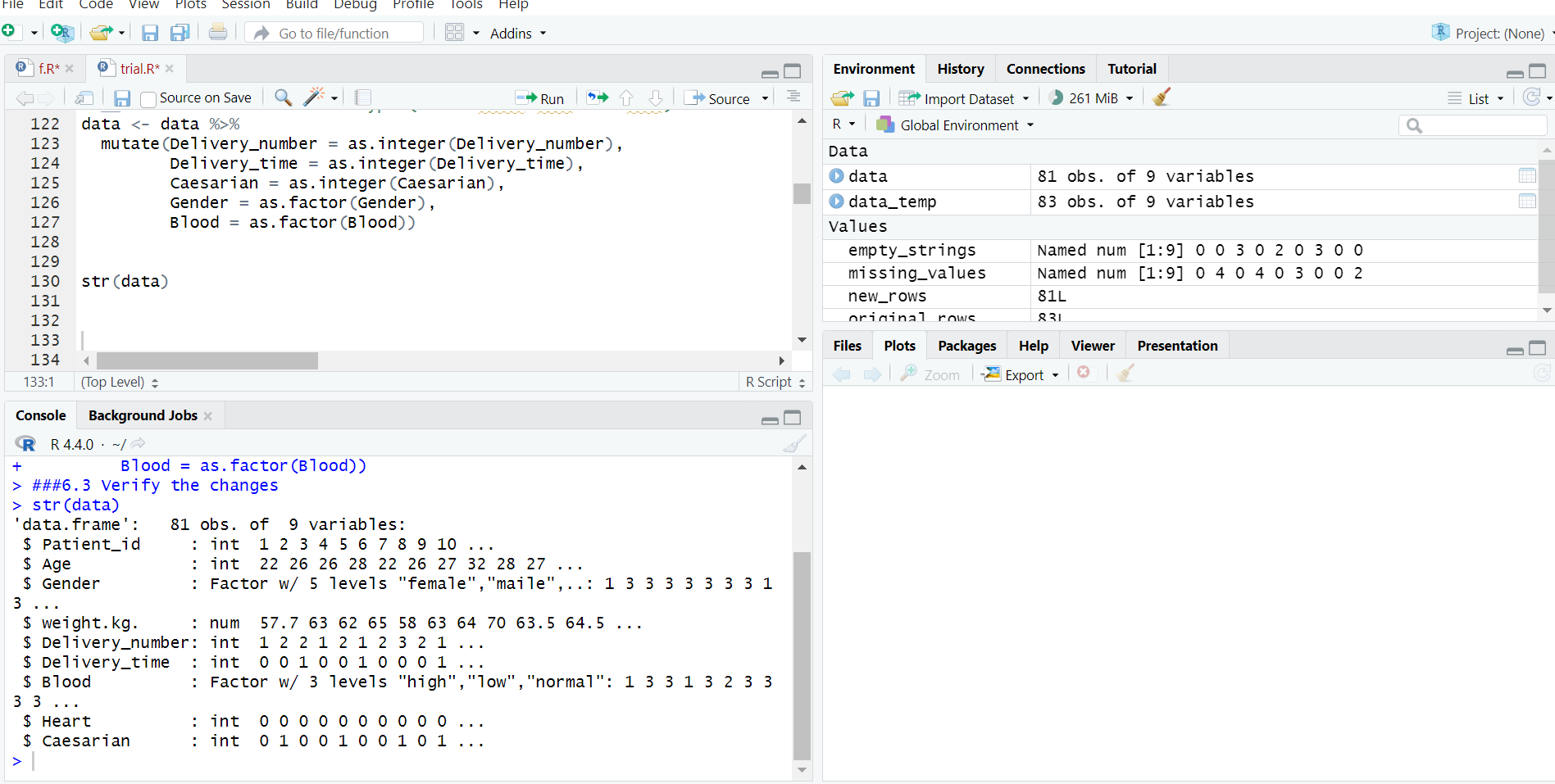
**Title: Verify the changes by checking the number of rows before and after removing duplicates**

1. **Convert data type check:**

This step involves examining the dataset's structure to ensure that each attribute is in the appropriate format required for analysis. It includes identifying the current data types (e.g., integers, floats, strings) and converting them as necessary. For example, numerical values that are mistakenly stored as strings might be converted to numeric types to facilitate statistical calculations. Additionally, continuous variables may be transformed into categorical types if required for specific analytical techniques. Proper data type conversion is essential to ensure the accuracy and reliability of data processing and subsequent analysis.



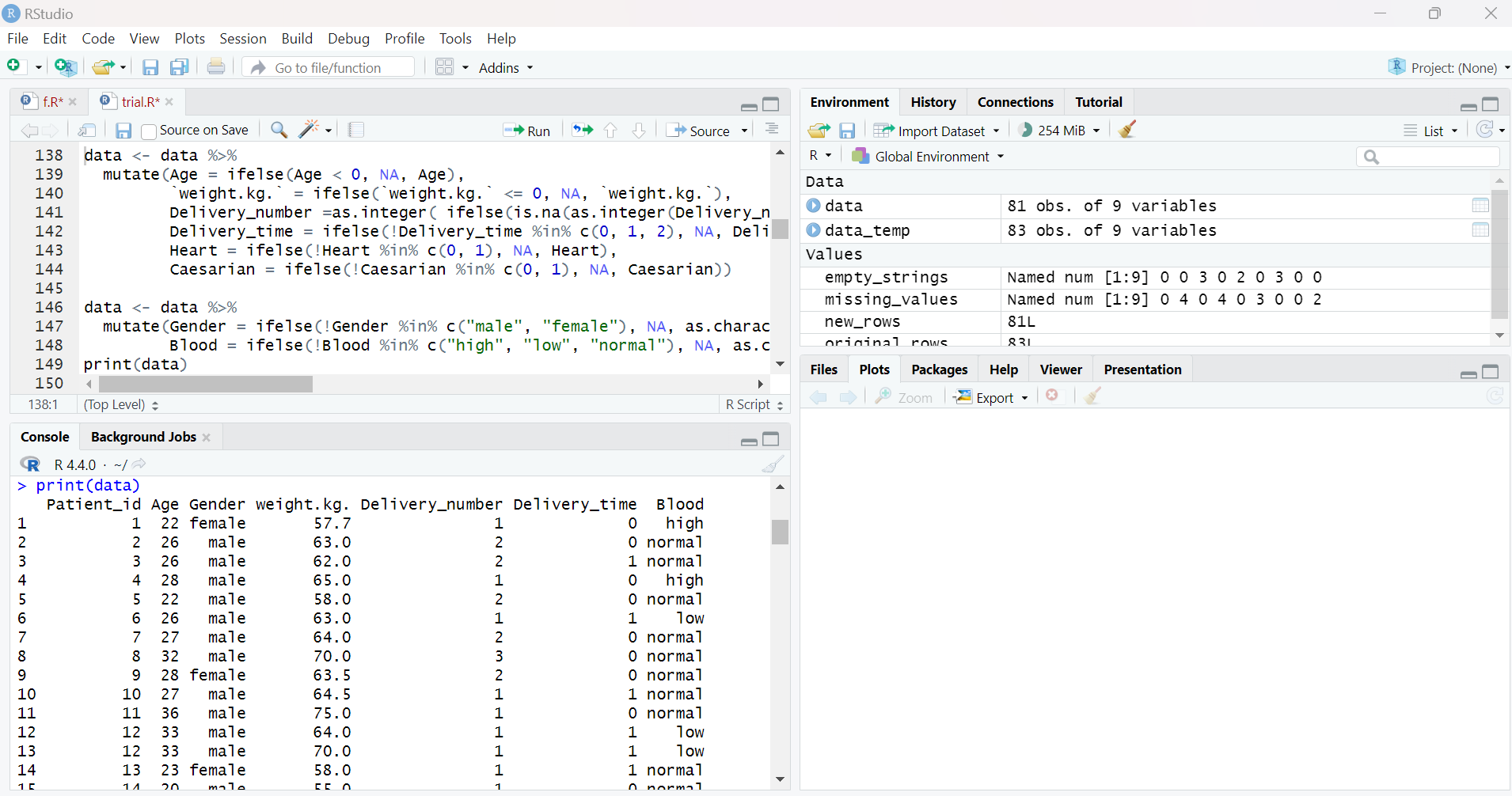
**Title: Inspect the structure of the dataset**



**Title: convert the data type**

1. **Handling invalid data**

Handling invalid data involves identifying and correcting or removing data points that are inaccurate, inconsistent, or fall outside the expected range. This process may include replacing invalid numeric values with NA (not available) to indicate missing or unprocessable data. After marking invalid data, imputation techniques can be applied, such as replacing NA values with the median or mean for numerical columns to maintain the dataset's integrity. Addressing invalid data is crucial for ensuring the reliability of data analysis and avoiding misleading results.



**Title: Replace any remaining invalid numeric data with NA and then impute**

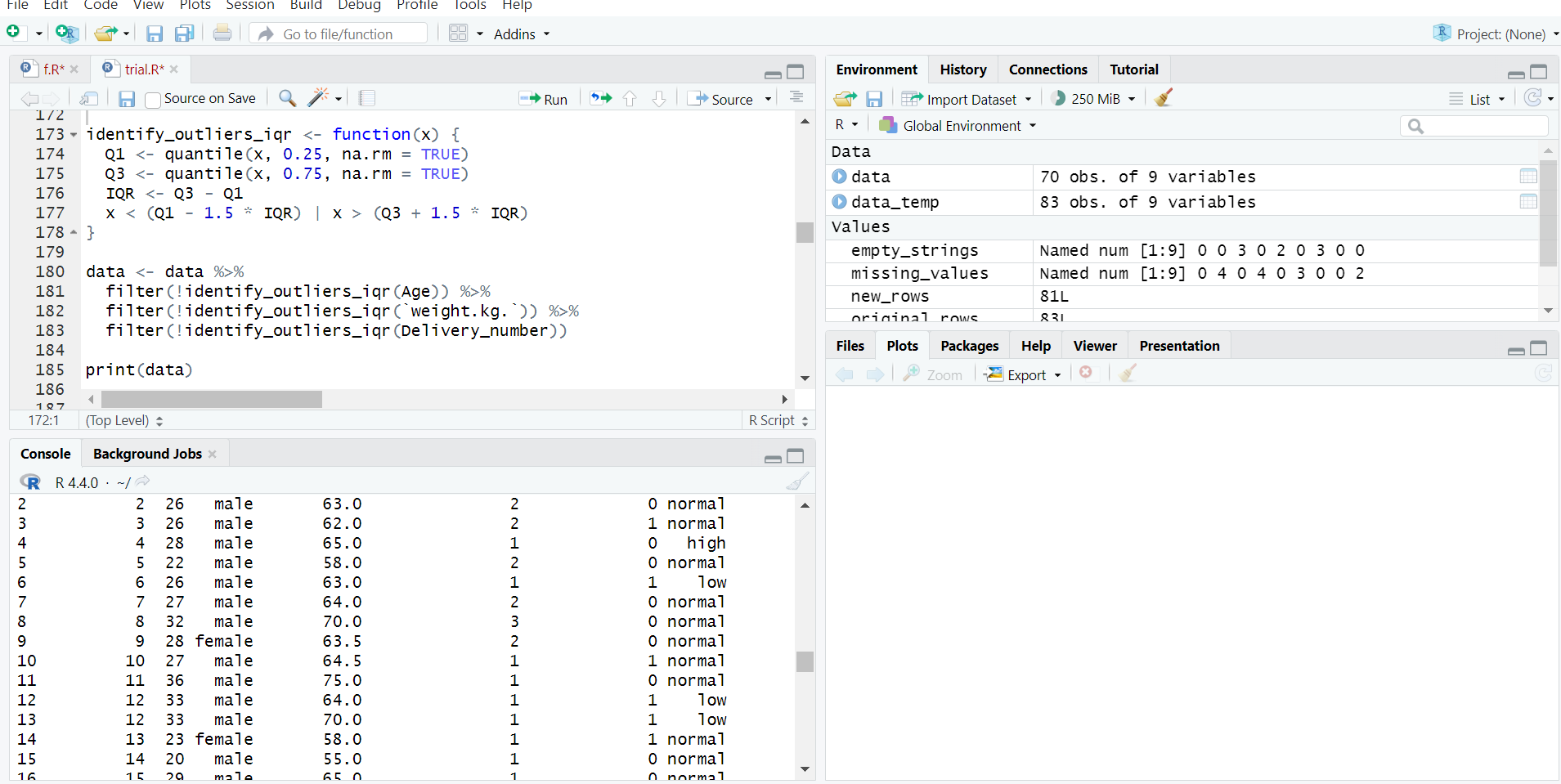
A screenshot of a computer

Description automatically generated

**Title: Replace NA with median for numerical columns**

1. **Outlier**

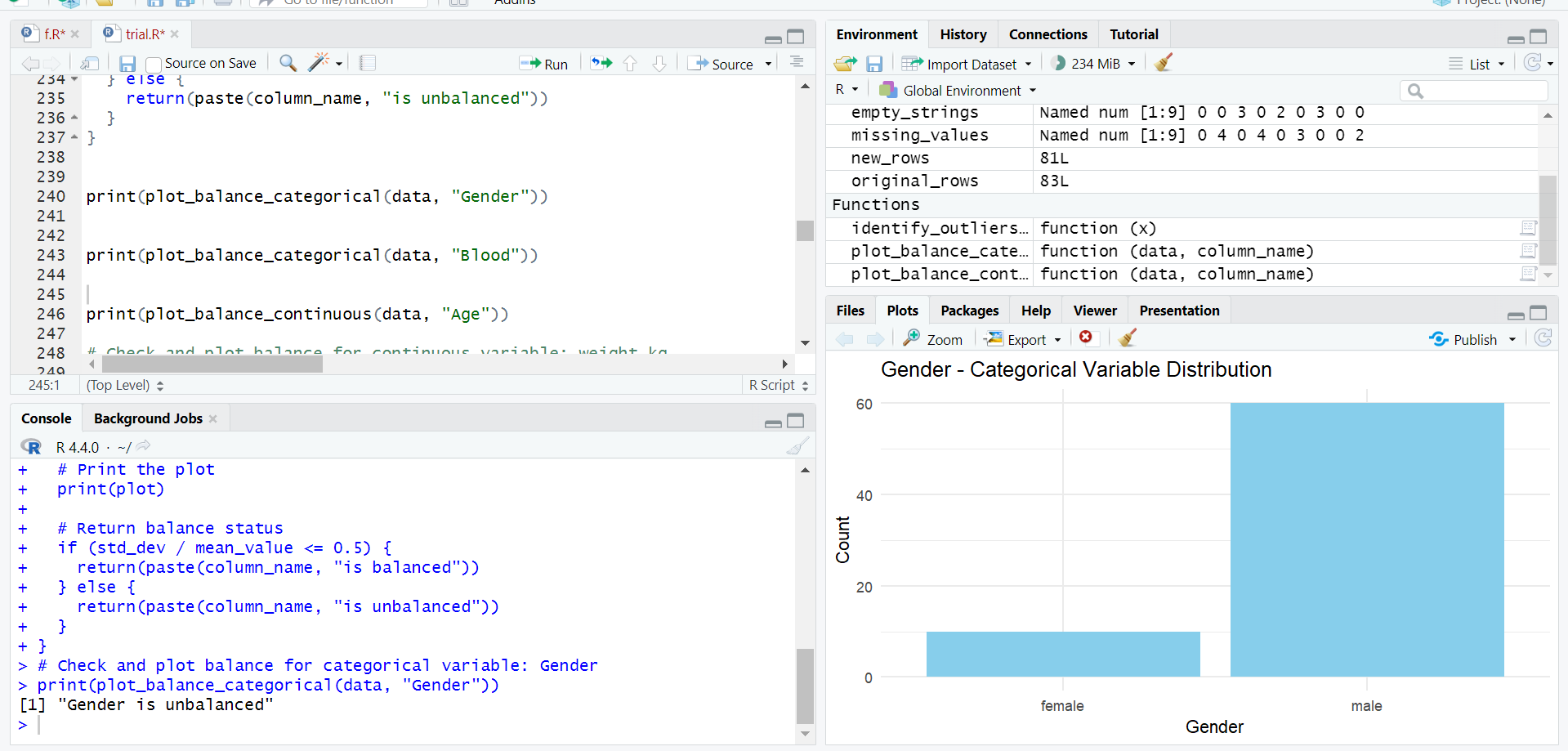
Outliers are data points that significantly differ from other observations in the dataset and can potentially distort the results of data analysis. This slide discusses identifying outliers using statistical methods like the Interquartile Range (IQR), which helps detect extreme values by calculating the range within which the central 50% of the data lies. Once identified, outliers can be handled by either removing them or transforming the data to minimize their impact. Properly managing outliers is essential to ensure the accuracy and validity of the analytical results.



**Title: Function to identify outliers using IQR and Identify and remove outliers using IQR**

1. **Balanced the dataset**

Balancing the dataset involves ensuring that different categories or classes within the data have a similar number of observations. An imbalanced dataset, where some classes have significantly more instances than others, can lead to biased models and inaccurate predictions. Techniques such as oversampling (increasing the number of instances in underrepresented classes), under sampling (reducing the number of instances in overrepresented classes), or generating synthetic data can be used to achieve balance. This step is crucial for improving the performance and reliability of machine learning models and ensuring that predictions are not skewed toward the majority class.



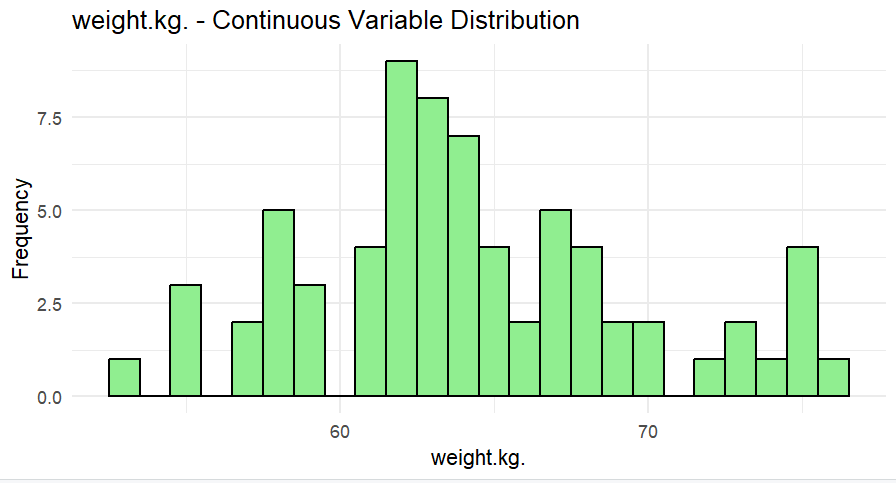
**Title: First checking the dataset which attributes are unbalanced**

A graph of a bar graph

Description automatically generated with medium confidence

A graph of a number of variable distribution

Description automatically generated



A graph of a number

Description automatically generated with medium confidence

A bar graph with blue squares

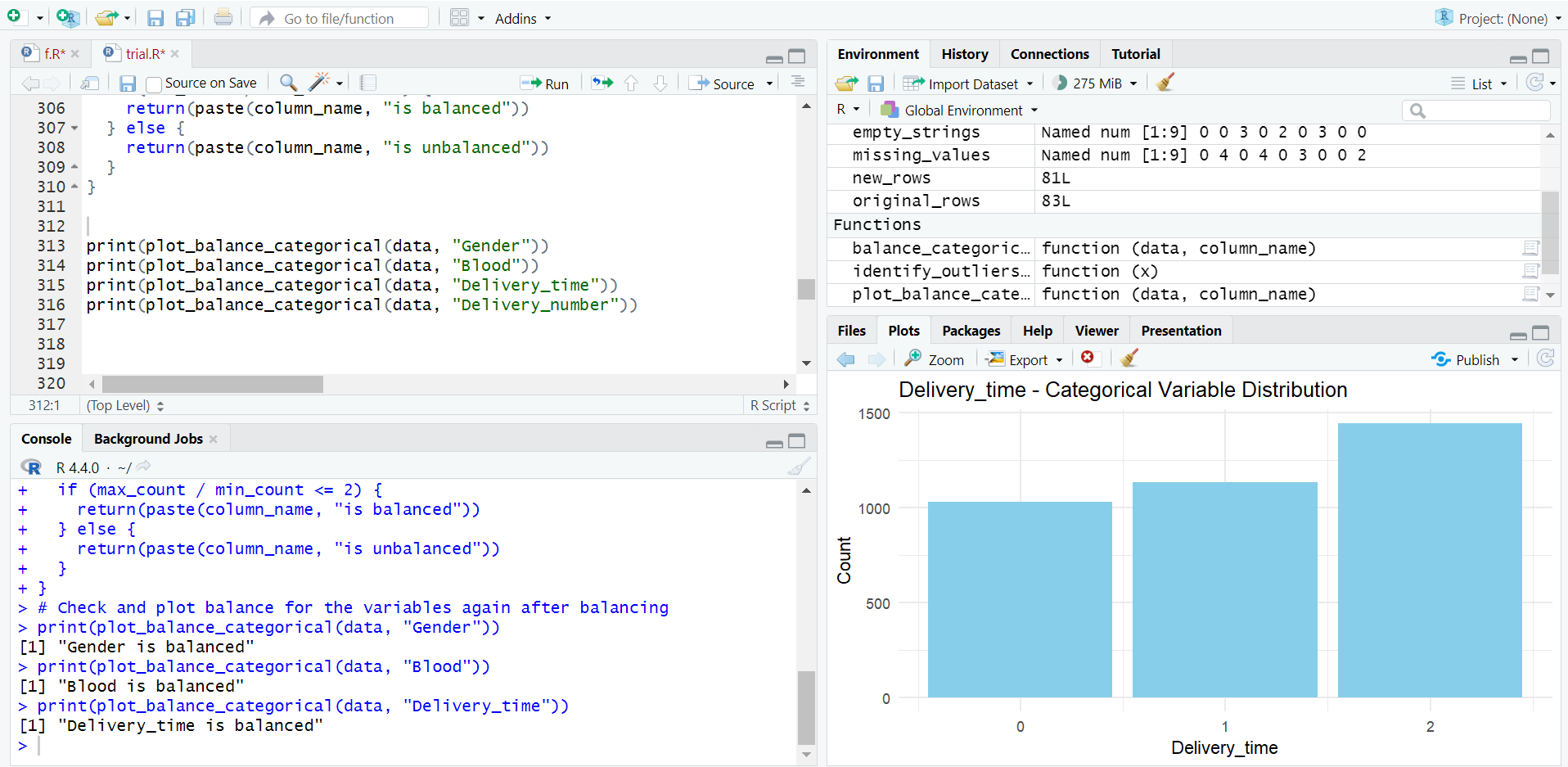
Description automatically generated

A graph with a bar graph

Description automatically generated

A graph of a bar graph

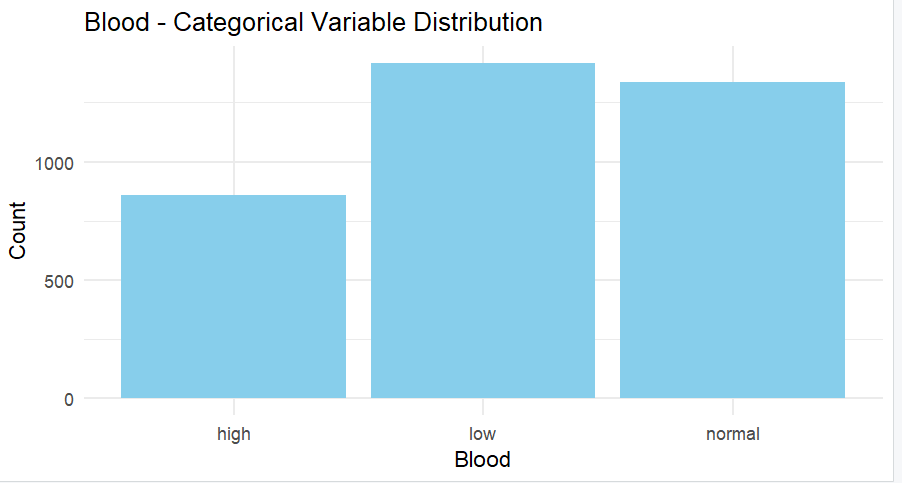
Description automatically generated with medium confidence



**Title: Balanced those unbalanced variable**

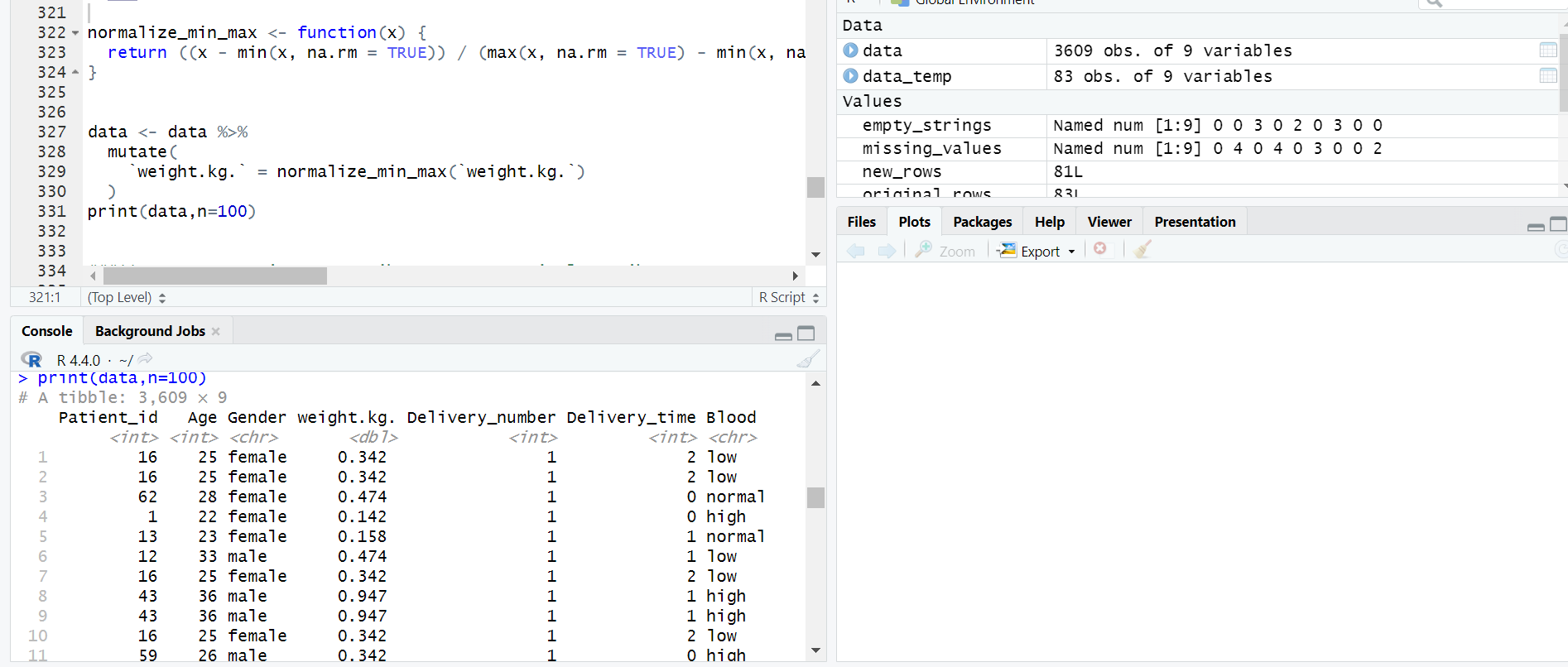
A screen shot of a graph

Description automatically generated



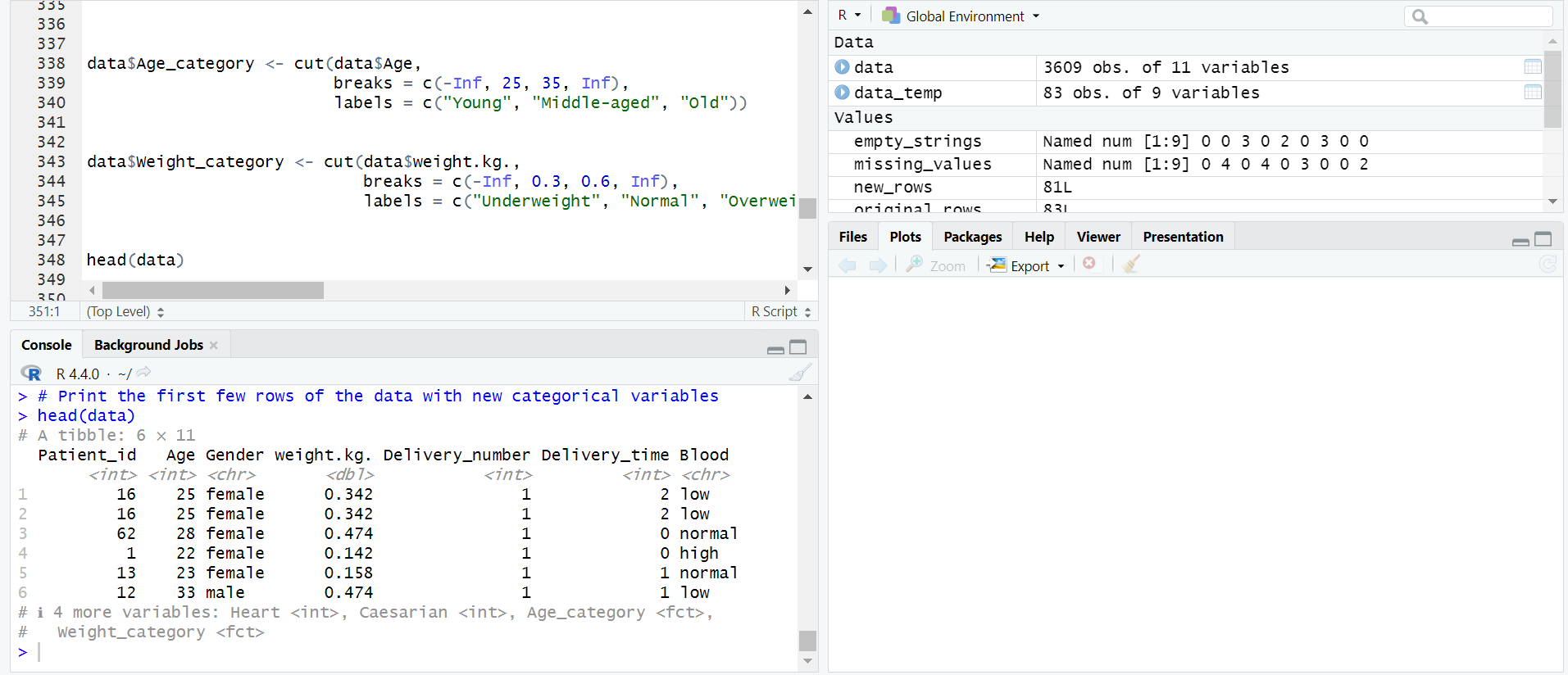
1. **Normalization**

Normalization is the process of scaling numerical data to a common range, typically between 0 and 1, without distorting differences in the ranges of values. This technique is essential when the dataset contains features with varying units or scales, as it ensures that no single feature dominates others due to its magnitude. By applying normalization, each feature contributes equally to the model's learning process, leading to more stable and efficient training, especially in machine learning algorithms sensitive to the scale of input data.



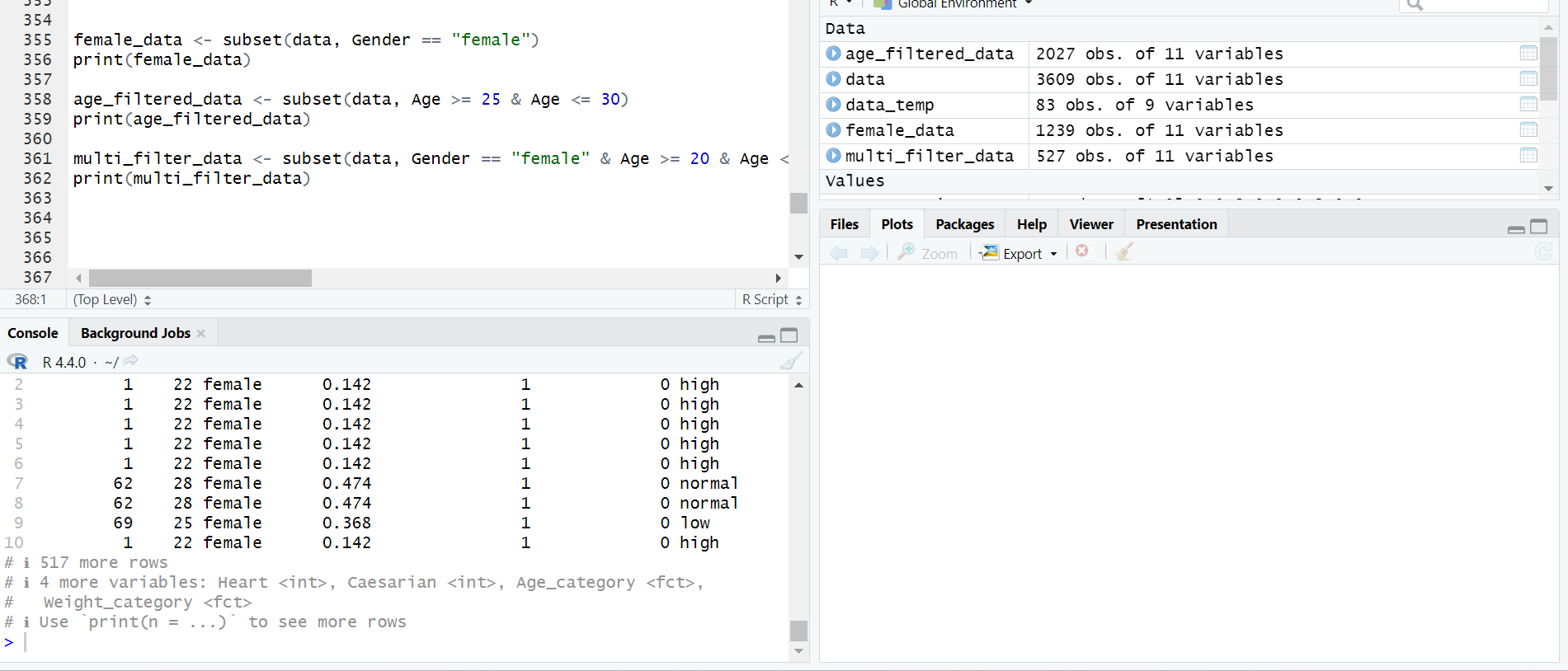
**11.Convert continuous attribute to categorical attributes**

Transforming continuous numerical variables into categorical ones. This transformation can be useful when numerical data need to be represented in distinct categories (e.g., age groups) for certain types of analysis or modeling.



**12.Filter the data**

The process of filtering the dataset to select specific data points that meet certain criteria. Filtering can be used to focus the analysis on relevant subsets of data, such as specific age groups, medical conditions, or other key characteristics.



**13. Find mean, median, mode with graph**

Calculating basic statistical measures—mean (average), median (middle value), and mode (most common value)—and visualizing them with graphs. These measures provide insights into the central tendency of the dataset, helping understand the distribution of data values.



A graph with numbers and lines

Description automatically generated

A graph with blue squares

Description automatically generated