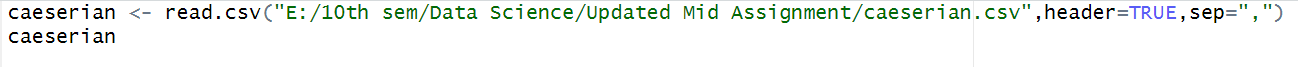
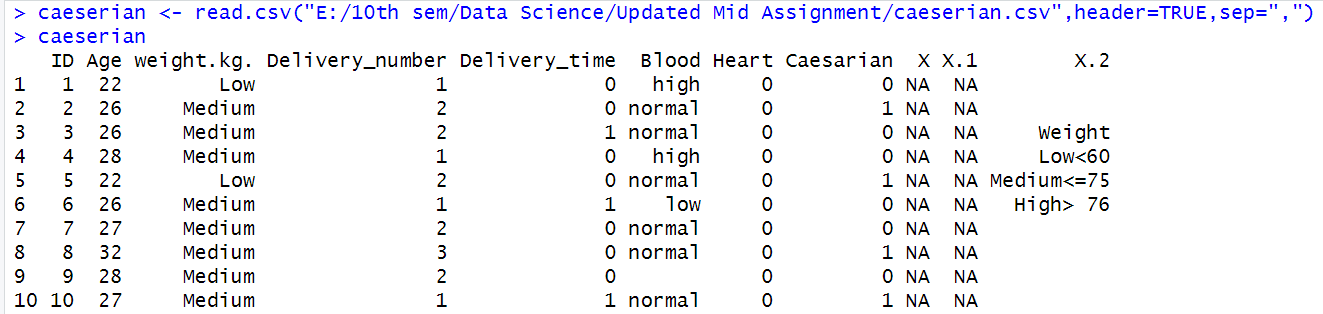
**Caesarian Section Classification Dataset:**

The Caesarian Section Classification Dataset from the UCI Repository contains 80 instances of medical records related to childbirth. Each record includes five attributes: age of the mother, number of previous deliveries, delivery time (timely or not), blood pressure (high or normal), and presence of a heart problem. The target variable indicates whether a Caesarian section was recommended (1) or not (0). The dataset is primarily used for binary classification tasks in machine learning, especially in the healthcare domain. It provides a practical resource for exploring data preprocessing, statistical analysis, and classification techniques in medical decision-making scenarios.

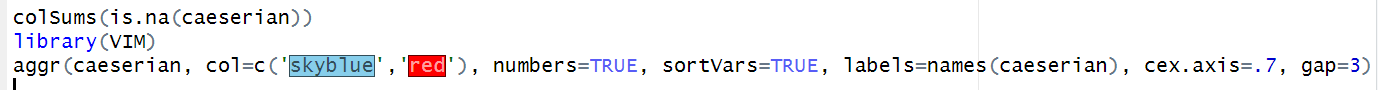
**Load Dataset:**

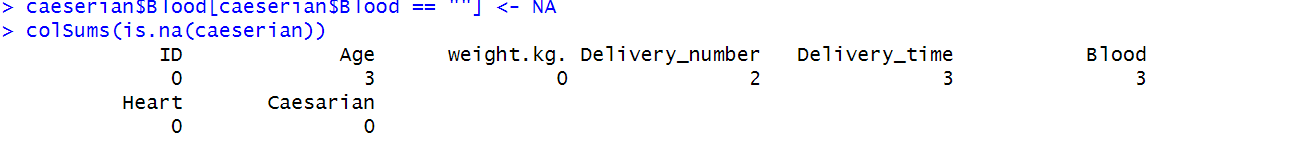
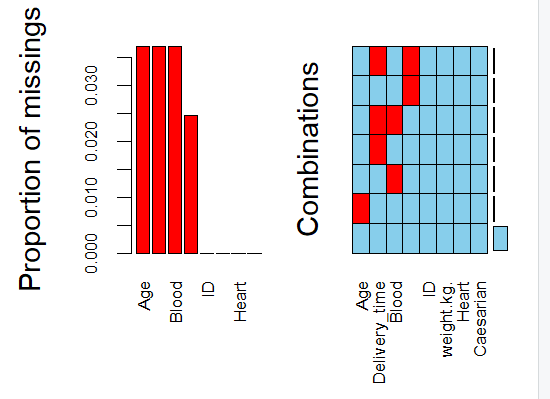




In R program for loading csv file we use read.csv.

**Missing Values in graph:**

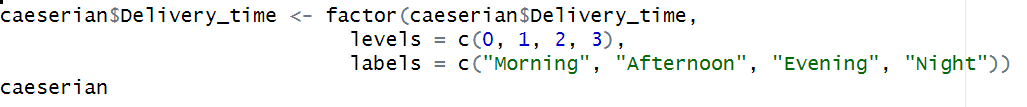


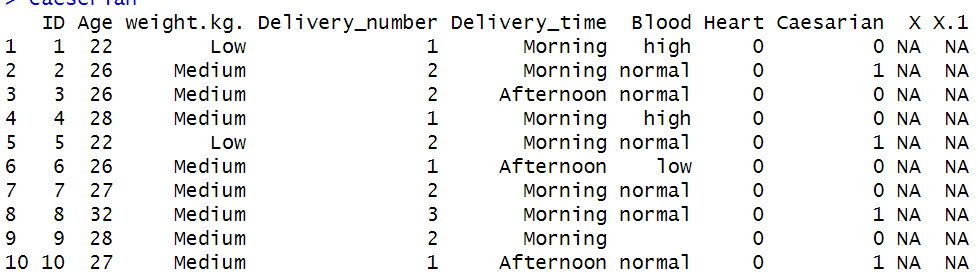


The dataset contains missing values in Age, Delivery\_number, Delivery\_time, and Blood columns. Using the VIM package, a graphical visualization shows the proportion and combination of missing entries, helping us understand where imputation or removal is necessary during preprocessing.

**Converting Numerical to Categorical:**

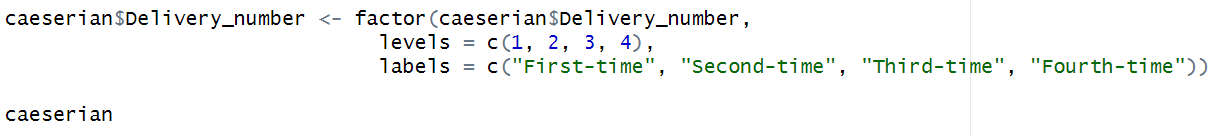
**Delivery\_time:**

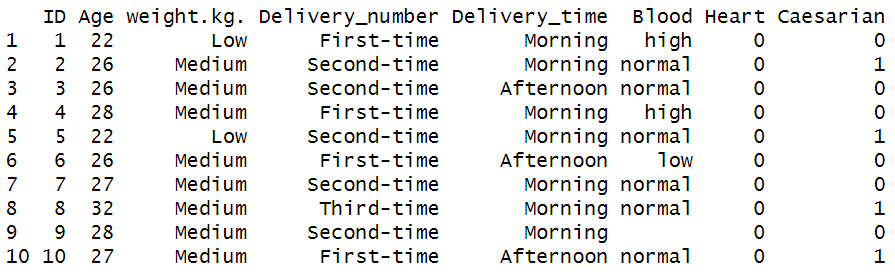




We need to convert Delivery\_time column data numeric to Categorical because this column represents the categorical data though it is numeric data.

**Delivery\_number:**

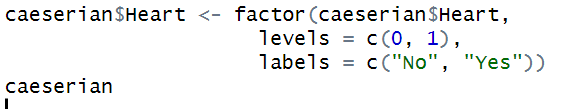


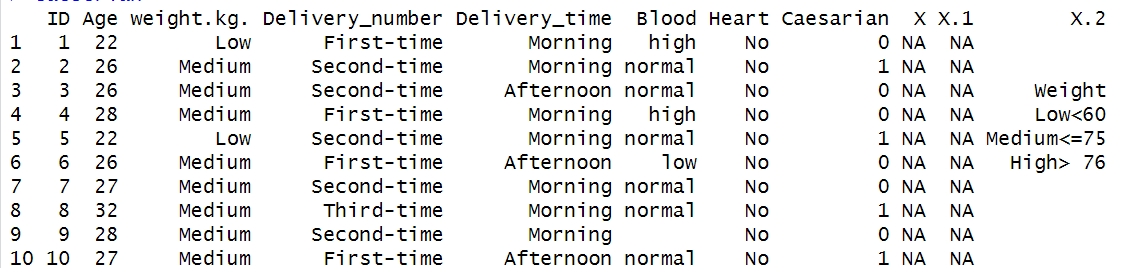


We convert the Delivery\_number columns numeric data to Categorical. For numeric number, 1=First\_time, 2=Second-time, 3=Third-time, 4=Fourth-time are labeled.

**Convert Categorical to Numeric:**

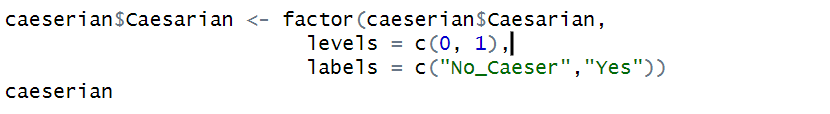
**Heart:**

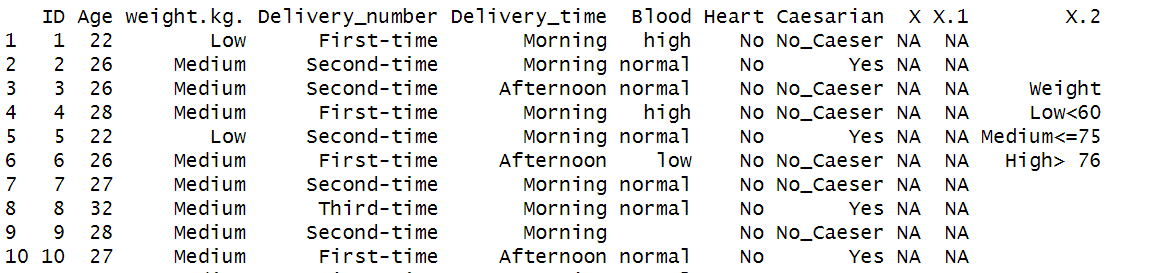




The Heart column was confirmed as a binary variable and converted into labeled factors: “Yes” for presence(1), “No” otherwise(0).

**Caesarian:**

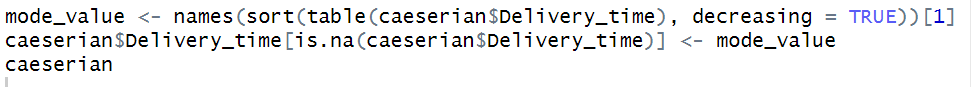


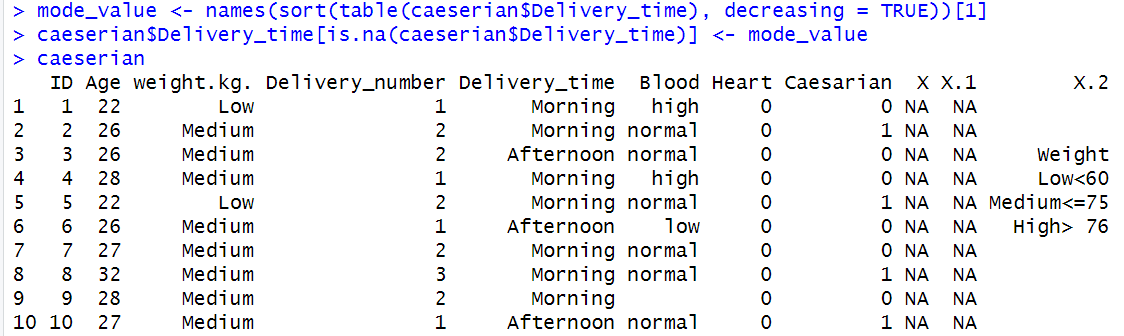


The Caesarian column indicates whether a caesarian delivery was performed. It was confirmed as a binary target variable and converted into categorical labels—“Yes” for caesarian(1) and “No-Caseser” for normal delivery(0) —for improved clarity and classification modeling.

**Handle Missing Values:**

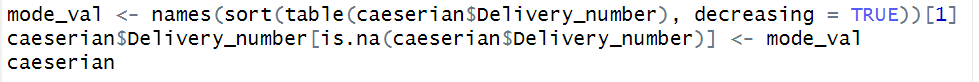
**Delivery\_time:**

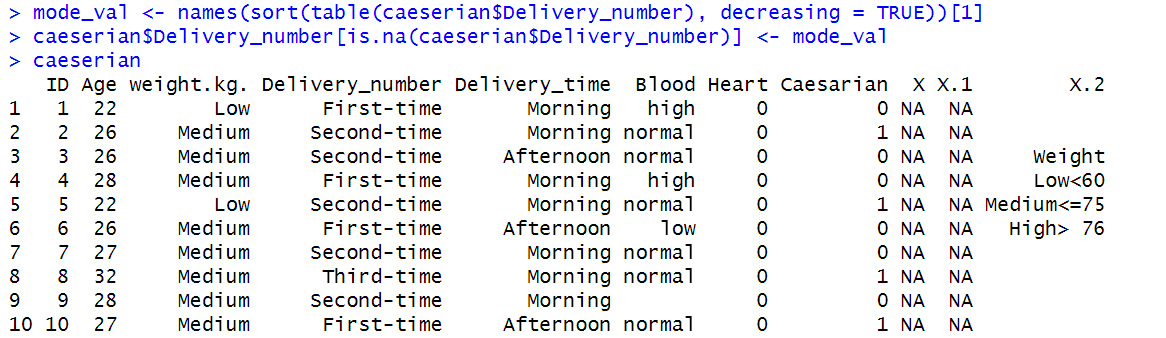




Missing values are handled using the most frequent category. This approach preserves the categorical structure and ensures data consistency, replacing NAs with the most commonly occurring label such as "Morning" or "Afternoon".

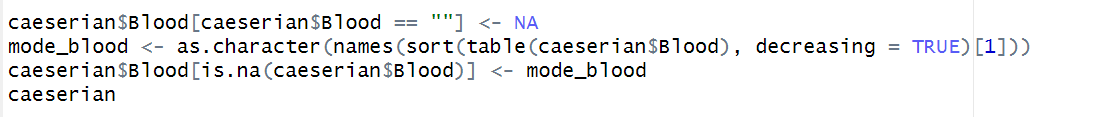
**Delivery\_number:**

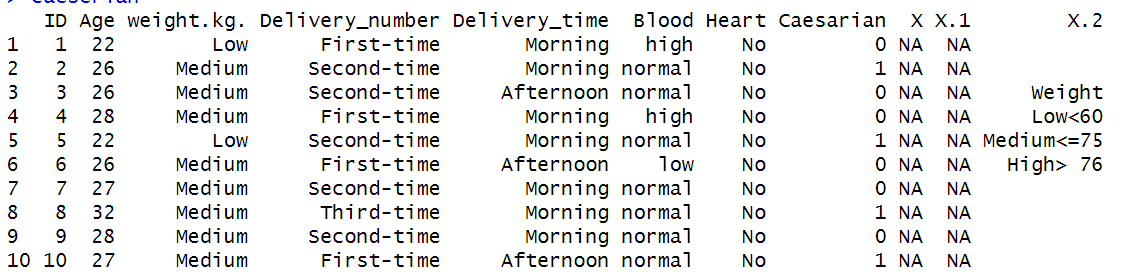




Missing values are handled using the most frequent category. This approach preserves the categorical structure and ensures data consistency, replacing NAs with the most commonly occurring label.

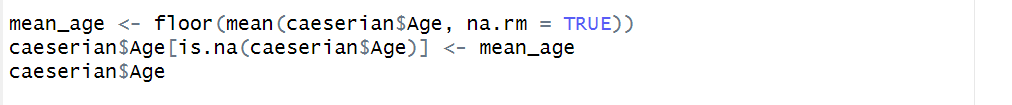
**Blood:**

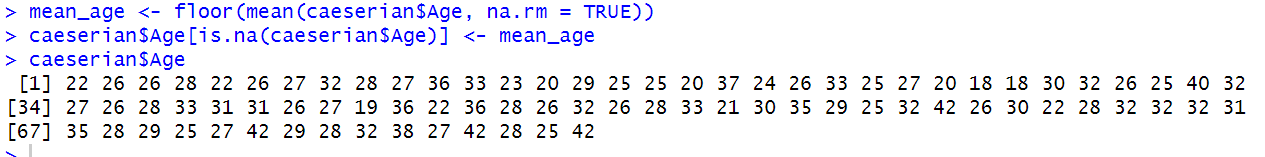




The Blood column contained blank entries which were not initially recognized as missing. These blanks were converted to NA, and all rows with missing Blood values were removed to ensure data quality and consistency in further analysis.

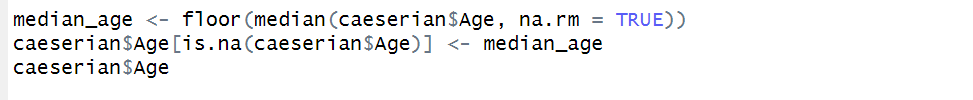
**Age: Mean()**

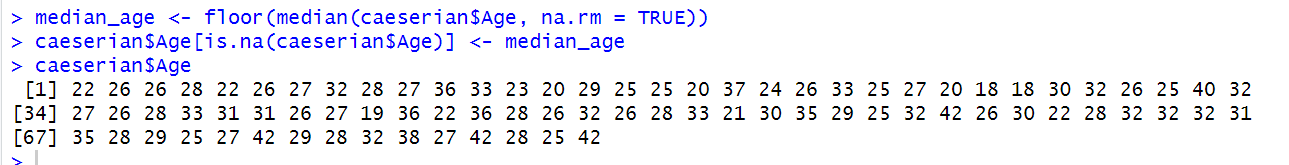




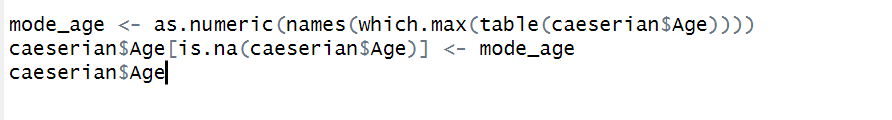
Missing values in the Age column were imputed using the mean. The mean value was calculated with na.rm = TRUE and then rounded down using floor() to maintain integer consistency. All NA entries were replaced with this floored mean value.

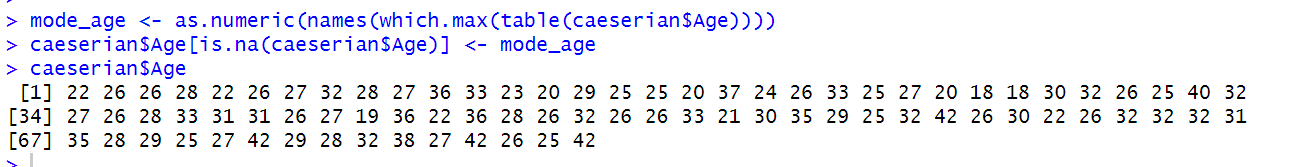
**Age: Median()**



Missing Age values were filled using the median. To maintain integer consistency, the median was rounded down using floor() before imputation, preserving the central tendency of the dataset.

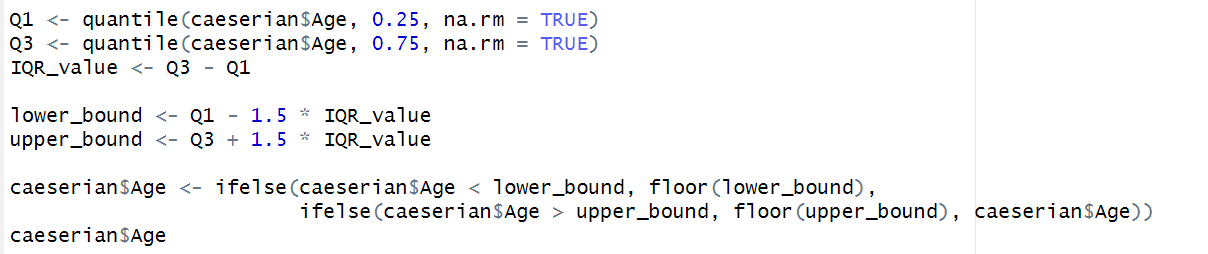
**Age: Mode()**

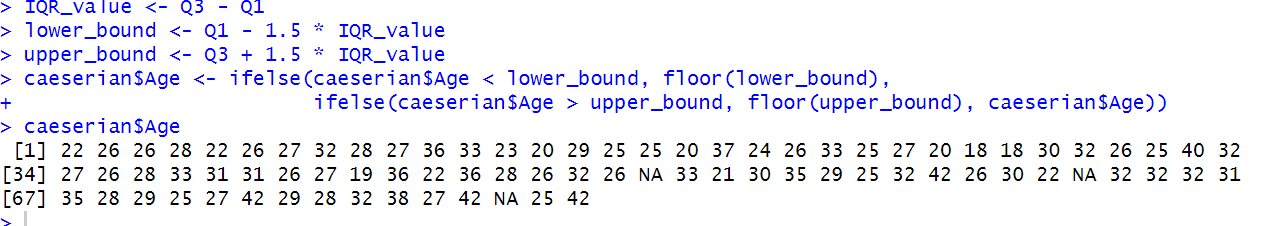




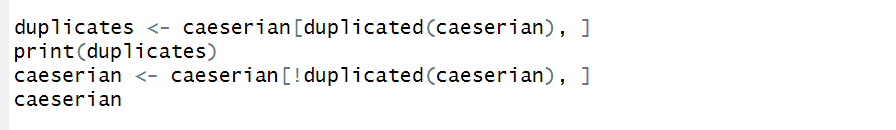
Missing Age values were filled using the most frequent value (mode), ensuring consistency with existing data. As mode is an integer, no rounding was necessary during the imputation process.

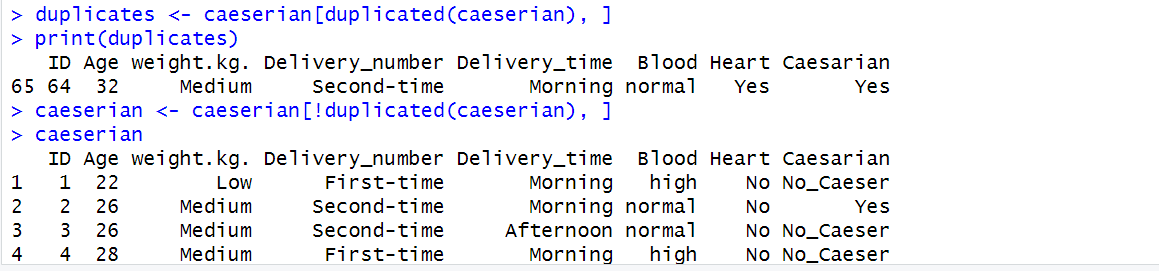
**Outlier Handle:**



  
Outliers in the Age column were capped within IQR-based limits using floor() to round down boundary values. Extreme ages below or above the acceptable range were replaced with the floored lower or upper bound, preserving all data without deletion.

**Find and Remove Duplicate:**

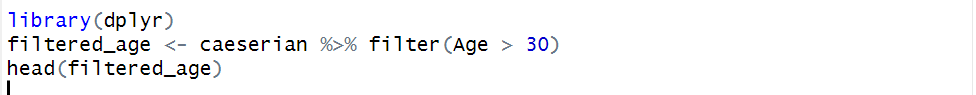


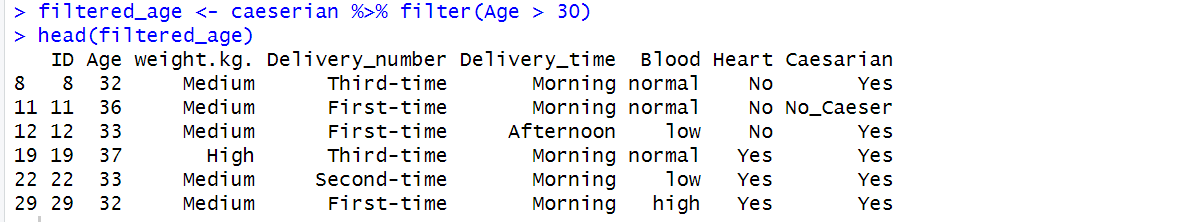


Two duplicate records with identical values (Age = 32, Weight = Medium, Delivery\_time = Morning, Heart = Yes) were identified and removed using duplicated(). This ensured uniqueness and prevented redundancy in the dataset for accurate analysis.

**Filtering Data:**

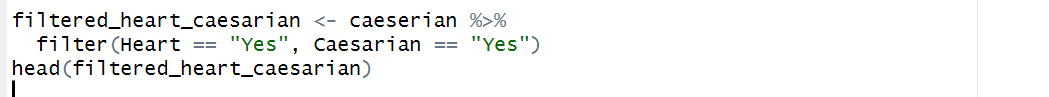
**Age > 30:**

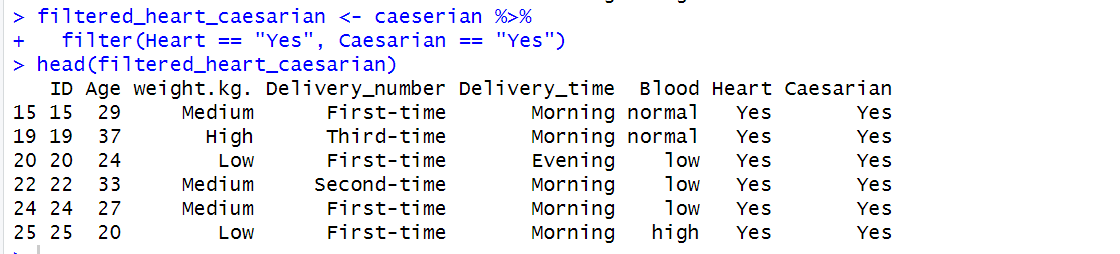




Patients aged above 30 were filtered using the filter(Age > 30) function. The resulting subset included individuals with varying delivery histories and health conditions.

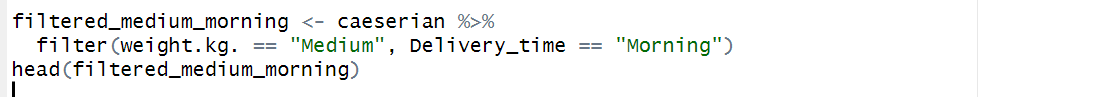
**Heart problem AND Caesarian**

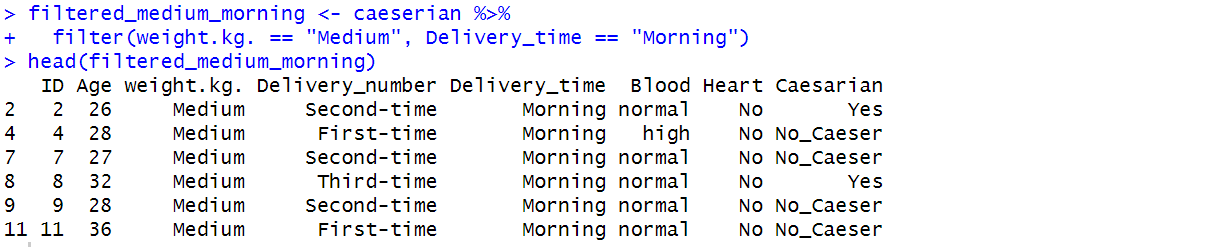




Patients with heart conditions who underwent Caesarian deliveries were filtered using filter(Heart == "Yes", Caesarian == "Yes"). The resulting subset included cases like ID 15 (Age 29, First-time delivery), enabling focused analysis of critical health profiles requiring surgical intervention.

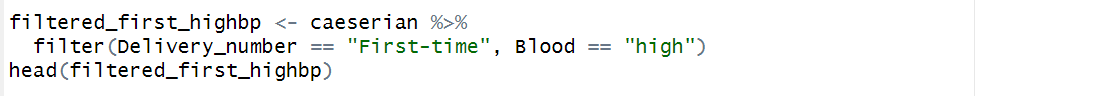
**Medium weight who delivered in Morning**

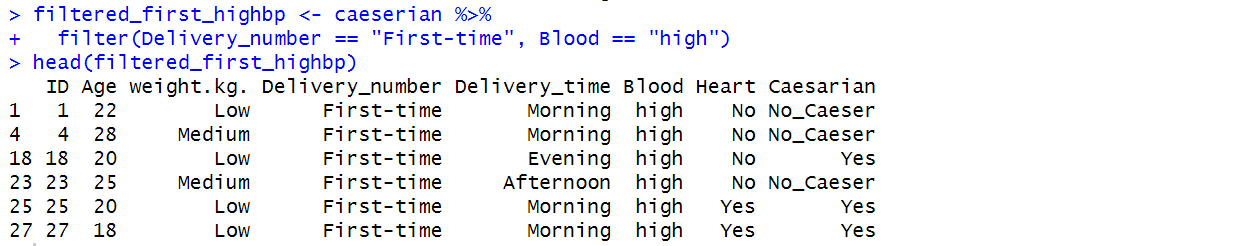




Patients with medium weight who delivered in the morning were filtered using filter(weight.kg. == "Medium", Delivery\_time == "Morning"). The subset includes cases like ID 2 (Age 26, Second-time), supporting analysis of delivery patterns by weight and time.

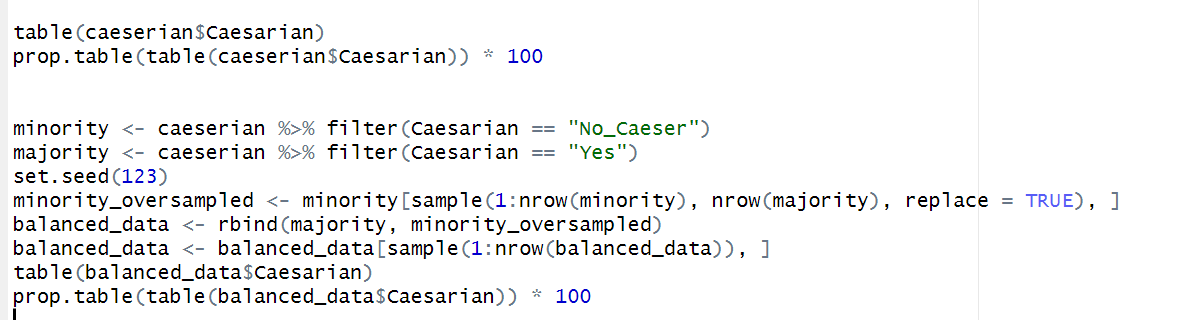
**First-time deliveries with High blood pressure**

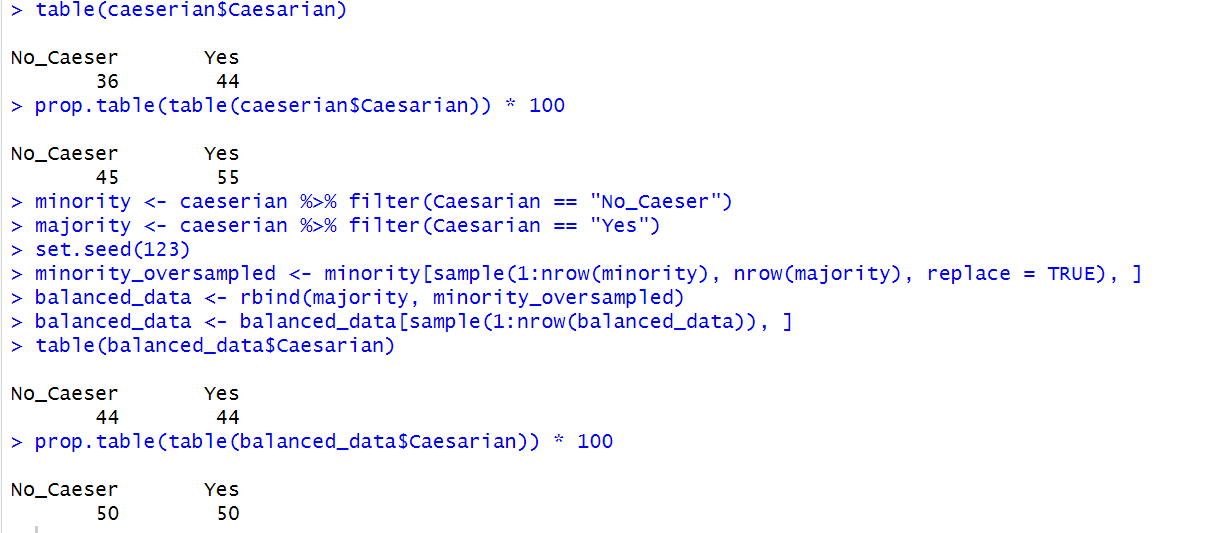




Patients undergoing first-time deliveries with high blood pressure were filtered using filter(Delivery\_number == "First-time", Blood == "high"). The subset includes cases like ID 1 (Age 22) and ID 27 (Age 18), useful for analyzing hypertension-related delivery outcomes.

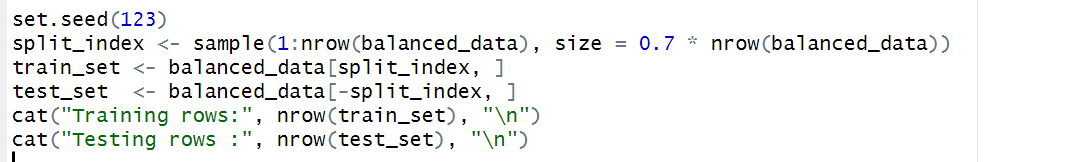
**Deal With Balanced and Imbalanced**

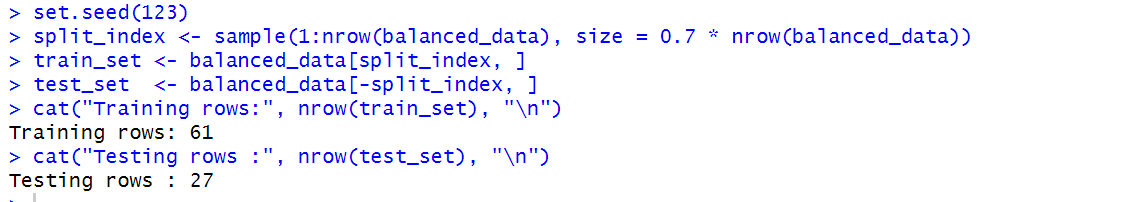




Initially, the dataset showed a slight imbalance: 39 "Yes" and 34 "No\_Caeser" cases. The minority class was identified programmatically and oversampled using replacement to match the majority, ensuring balanced class representation without removing any original data.

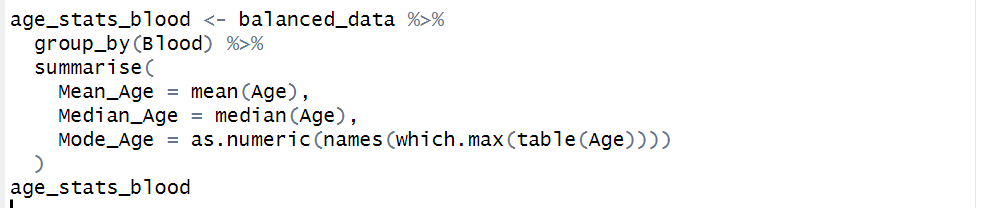
**Train Test Split:**

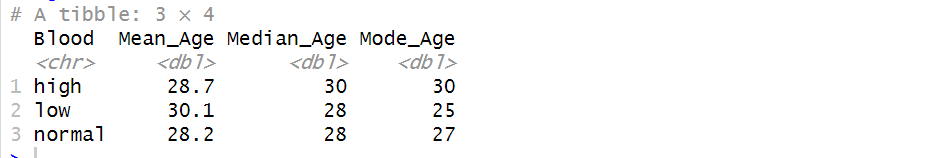




The final balanced dataset was randomly split into training (70%) and testing (30%) subsets using sample(). This split ensures unbiased model evaluation.

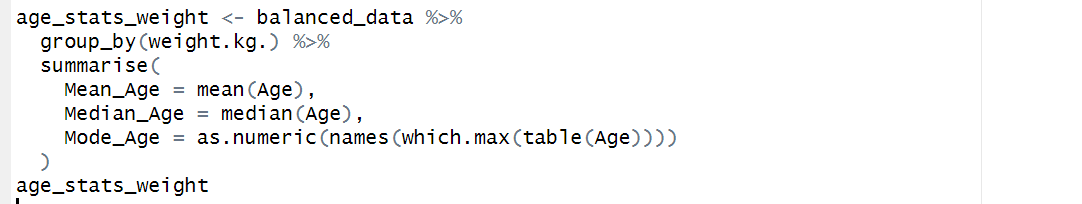
**Central tendencies (mean, median, mode) of Age across Blood**

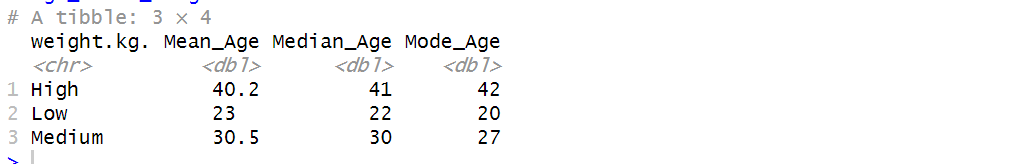




Patients with normal blood pressure showed the highest mean and median age. The mode varied slightly across groups, suggesting that age distributions differ subtly based on blood pressure levels.

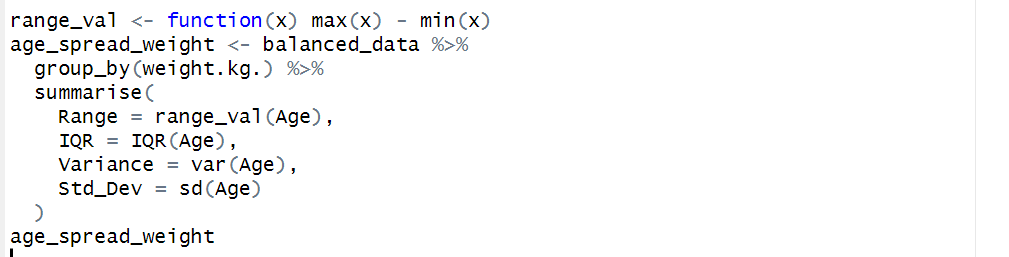
**Central tendencies (mean, median, mode) of Age across Weight**

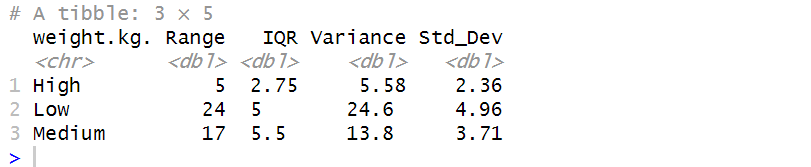




The Age distribution was compared across different weight categories. Patients in the High weight group showed the highest average and median age. The mode age also increased from Low to High, indicating a positive relationship between weight and maternal age.

Compare the Spread (Range, IQR, Variance, Standard Deviation) of Age across Weight





The spread of Age varied across weight groups. Medium-weight patients had the widest IQR and highest variance, indicating greater age variability. The standard deviation also peaked in this group, suggesting more dispersed age values compared to low and high weight groups.