

American International University- Bangladesh

Data Science

Final Project Report

Summer 22-23

Name : Nayeem Abdul Qaiyum

ID : 20-43581-1

Section : C

Date Of Submit: 15.08.23

Submitted By

# Abdus Salam

Assistant Professor, CS

Data Set Name: US Births 👶 by Year, State, and Education Level

**Data set Link:** https://www.kaggle.com/datasets/danbraswell/temporary-us-births

**Description:**

This dataset provides birth rates and related data across the 50 states and DC from 2016 to 2021. A particular emphasis is given to detailed information on the mother's educational level.

There are several rows and 9 columns in the data set and they are – State, State.Abbreviation, Year, Gender, Education.Level.of.Mother, Education.Level.Code, Number.of.Births, AverageAge.of.Mother..years., Average.Birth.Weight..g. There are different types of attributes in this dataset and they are integer, numeric, character. Here we apply KNN method to find the highly accurate results.

Table of Contents

* Import data
* View the structure of the dataset
* Column name of the data set
* Summary
* Data preparation steps
* Conversion
* Missing Value
* Normalization
* Correlation
* Delete Column
* Plot Correlation Matrix
* Training & Testing
* Accuracy
* 10-fold cross validation
* Confusion matrix

**Project Solution**

**Import data:**

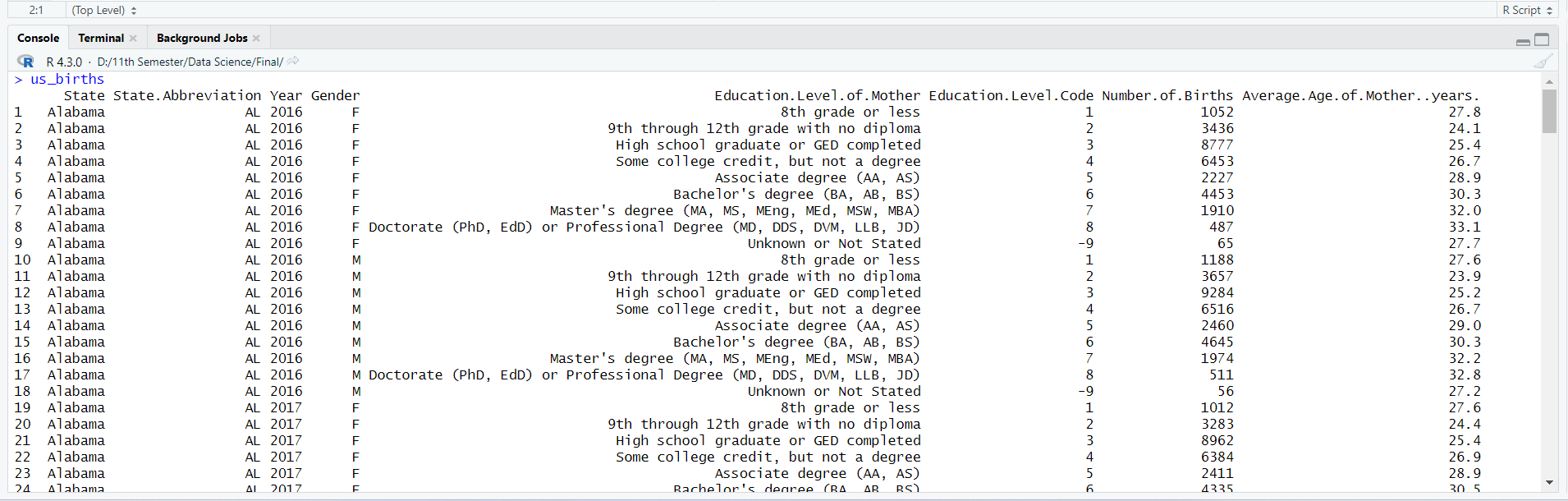
Insert all of the data from the excel file first, and then save the document as a dataset file. then convert the dataset file's format to a CSV file. After importing my CSV file into RStudio, I add the following code.

**Code Segment:**

us\_births <- read.csv("D:/11th Semester/Data Science/Final/us\_births\_2016\_2021.csv")

us\_births

**Output:**

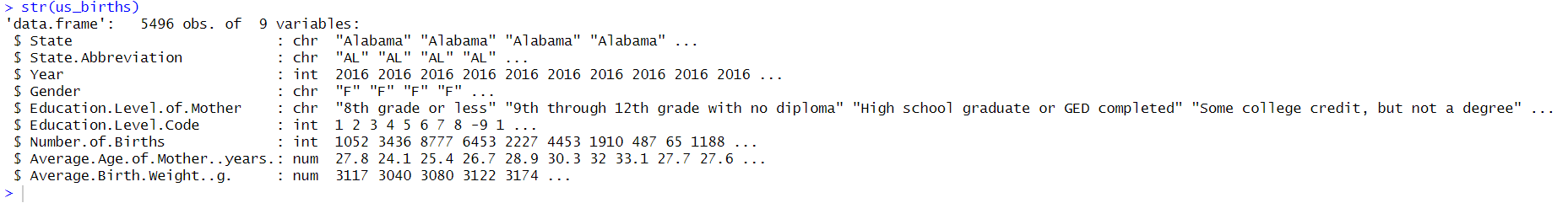
****

**View the structure of the dataset:**

The dataset structure is shown using the str() function, including the variables, their data types, and the initial values. We will get a general idea of the dataset from this.

**Code Segment:**

str(us\_births)

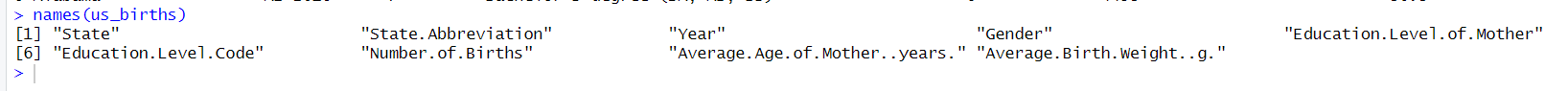


**Column name of the data set:**

**Explanation:** To see the all column name we using the names() function.

**Code Segment:**

names(us\_births)

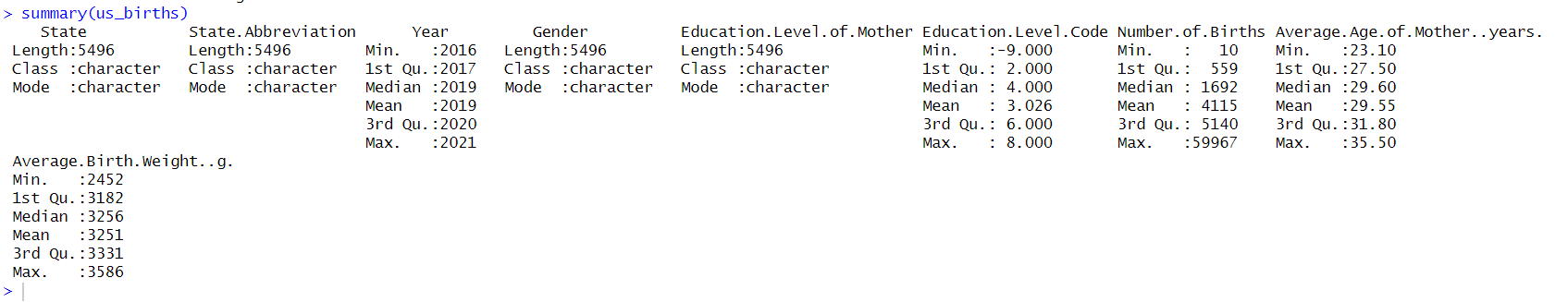
****

**Summary:**

For numerical variables in the dataset, the summary() function returns summary statistics (count, mean, median, etc.). This will help us gain understanding of the variables' distribution and central patterns.

**Code Segment:**

summary(us\_births)



**Data preparation steps**

First, I need to prepare my dataset so that I can apply the KNN method later.

To prepare my dataset firstly I need to convert all categorical data to numerical data. Also, we can delete any column unless we need it.

In this dataset I delete one column and that is State Abbreviation.

**Conversion**

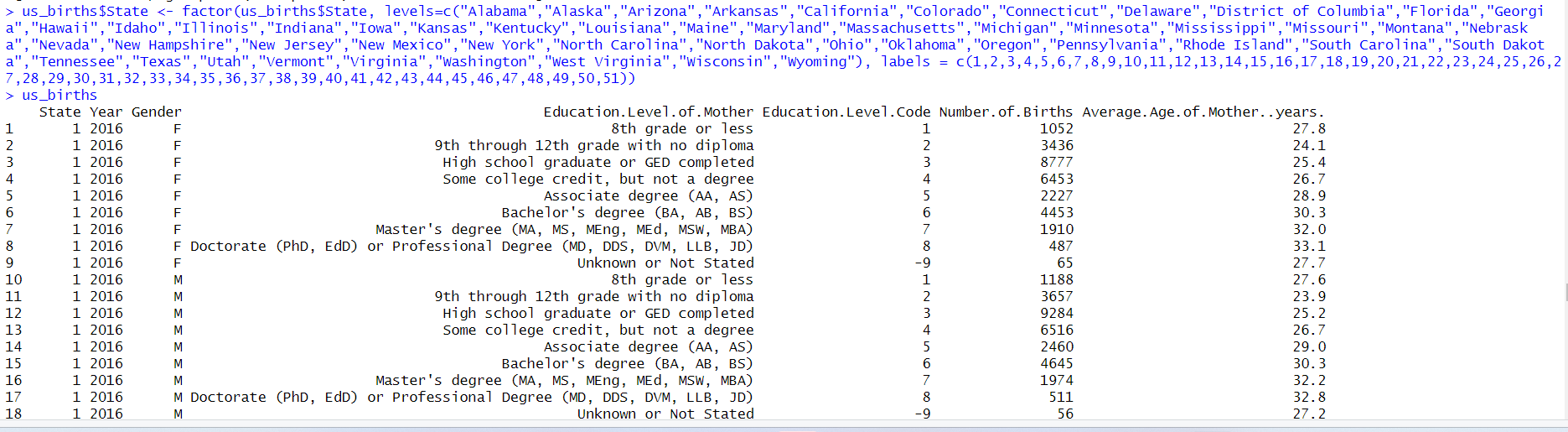
Converting categorical data to numerical data is a common preprocessing step in data science and analysis. This is often necessary because many algorithms, including K-Nearest Neighbors (KNN), work with numerical data and mathematical calculations.

**Categorical to Numeric (State column):**

**Code Segment:**

us\_births$State<factor(us\_births$State,levels=c("Alabama","Alaska","Arizona","Arkansas","California","Colorado","Connecticut","Delaware","District of Columbia","Florida","Georgia","Hawaii","Idaho","Illinois","Indiana","Iowa","Kansas","Kentucky","Louisiana","Maine","Maryland","Massachusetts","Michigan","Minnesota","Mississippi","Missouri","Montana","Nebraska","Nevada","New Hampshire","New Jersey","New Mexico","New York","North Carolina","North Dakota","Ohio","Oklahoma","Oregon","Pennsylvania","Rhode Island","South Carolina","South Dakota","Tennessee","Texas","Utah","Vermont","Virginia","Washington","West Virginia","Wisconsin","Wyoming"), labels = c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51))

us\_births

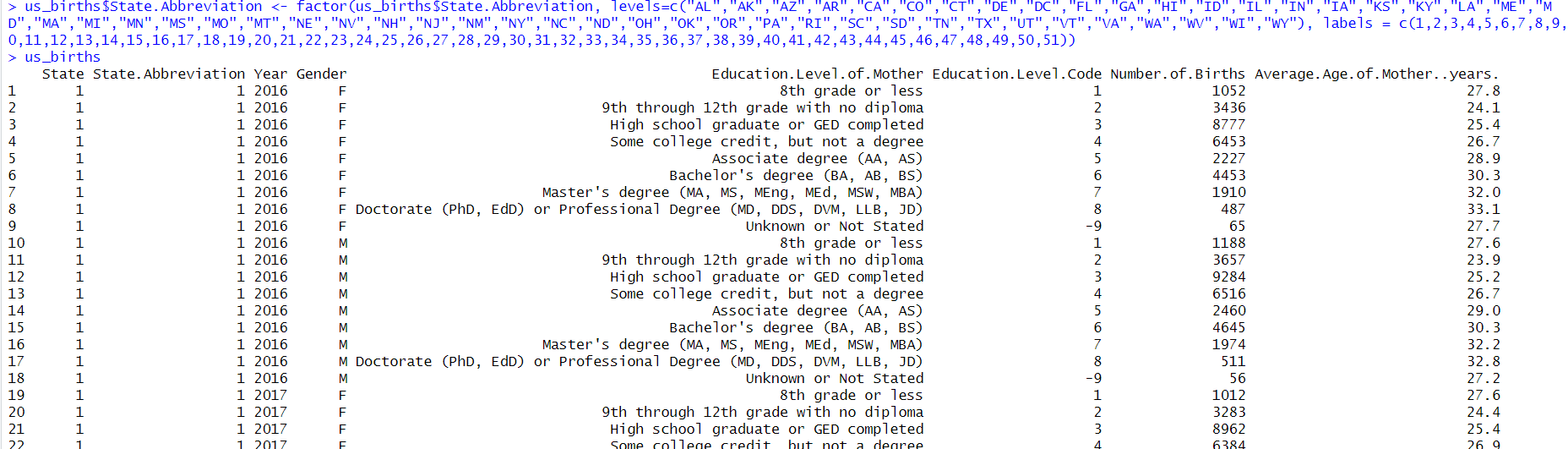


**Categorical to Numeric (State. Abbreviation column):**

**Code Segment:**

us\_births$State.Abbreviation <- factor(us\_births$State.Abbreviation, levels=c("AL","AK","AZ","AR","CA","CO","CT","DE","DC","FL","GA","HI","ID","IL","IN","IA","KS","KY","LA","ME","MD","MA","MI","MN","MS","MO","MT","NE","NV","NH","NJ","NM","NY","NC","ND","OH","OK","OR","PA","RI","SC","SD","TN","TX","UT","VT","VA","WA","WV","WI","WY"), labels = c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51))

us\_births

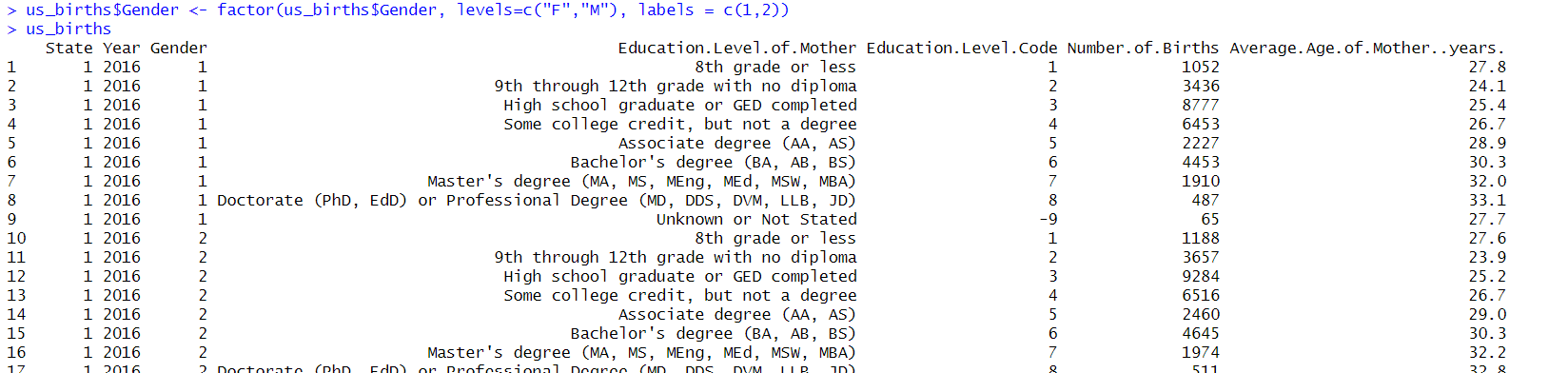


**Categorical to Numeric (Gender column):**

**Code Segment:**

us\_births$Gender <- factor(us\_births$Gender, levels=c("F","M"), labels = c(1,2))

us\_births

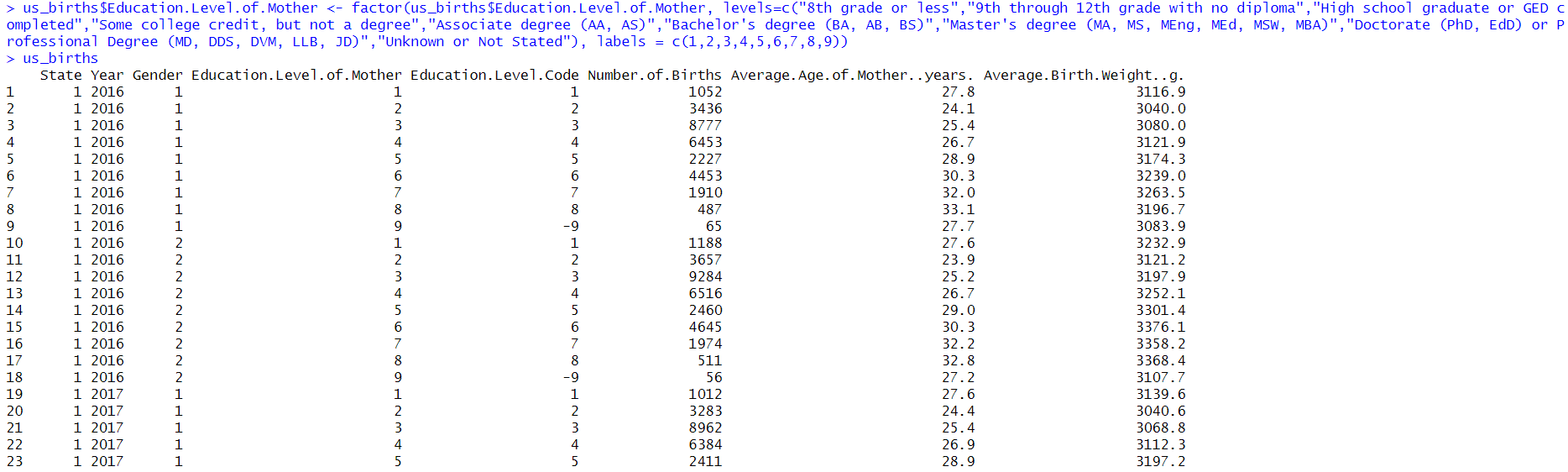


**Categorical to Numeric (Education Level of Mother column):**

**Code Segment:**

us\_births$Education.Level.of.Mother <- factor(us\_births$Education.Level.of.Mother, levels=c("8th grade or less","9th through 12th grade with no diploma","High school graduate or GED completed","Some college credit, but not a degree","Associate degree (AA, AS)","Bachelor's degree (BA, AB, BS)","Master's degree (MA, MS, MEng, MEd, MSW, MBA)","Doctorate (PhD, EdD) or Professional Degree (MD, DDS, DVM, LLB, JD)","Unknown or Not Stated"), labels = c(1,2,3,4,5,6,7,8,9))

us\_births



**Missing Value**

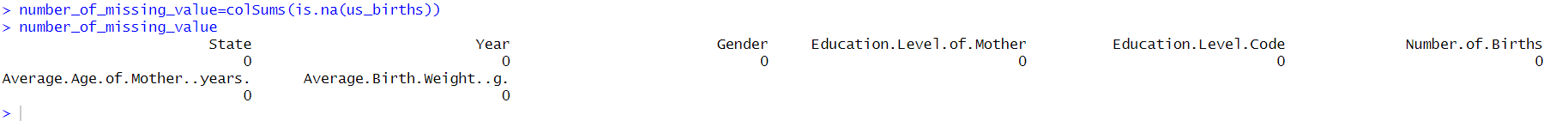
**Finding the missing value for all attributes:**

Missing data is crucial for accurate analysis and results.

**Code Segment:**

number\_of\_missing\_value=colSums(is.na(us\_births))

number\_of\_missing\_value



**Normalization**

Normalization is a data preprocessing technique that is commonly used in data science to scale and transform features to a consistent range (0,1). It involves adjusting the values of features in a dataset to ensure that they have similar scales.

**Code Segment:**

library(dplyr)

us\_births <- as.data.frame(sapply(us\_births, as.numeric))

min\_max\_norm <- function(x) {

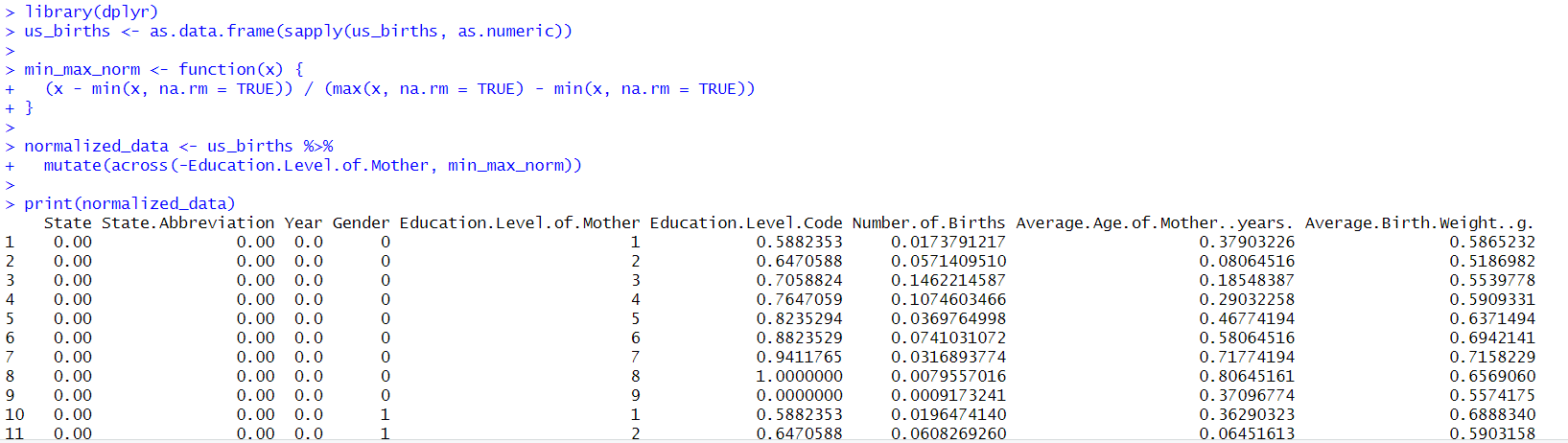
(x - min(x, na.rm = TRUE)) / (max(x, na.rm = TRUE) - min(x, na.rm = TRUE))

}

normalized\_data <- us\_births %>%

mutate(across(-Education.Level.of.Mother, min\_max\_norm))

print(normalized\_data)



**Correlation**

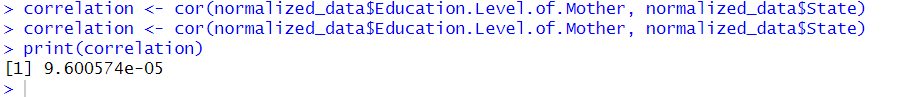
Correlation analysis is a statistical technique used to evaluate the strength and direction of the linear relationship between two or more variables in a dataset.

**Calculate the correlation between "Education.Level.of.Mother" and "State":**

**Code Segment:**

correlation <- cor(normalized\_data$Education.Level.of.Mother, normalized\_data$State)

print(correlation)

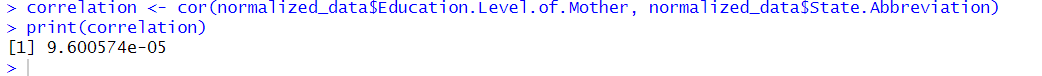


**Calculate the correlation between "Education.Level.of.Mother" and " State. Abbreviation":**

**Code Segment:**

correlation <- cor(normalized\_data$Education.Level.of.Mother, normalized\_data$State.Abbreviation)

print(correlation)

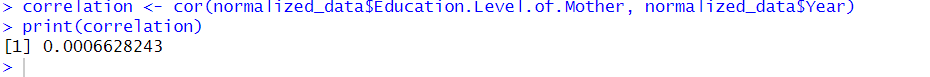


**Calculate the correlation between "Education.Level.of.Mother" and "Year":**

**Code Segment:**

correlation <- cor(normalized\_data$Education.Level.of.Mother, normalized\_data$Year)

print(correlation)

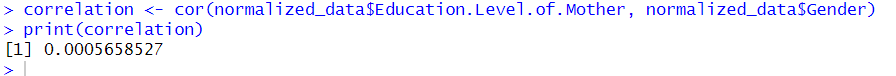
****

**Calculate the correlation between "Education.Level.of.Mother" and "Gender":**

**Code Segment:**

correlation <- cor(normalized\_data$Education.Level.of.Mother, normalized\_data$Gender)

print(correlation)

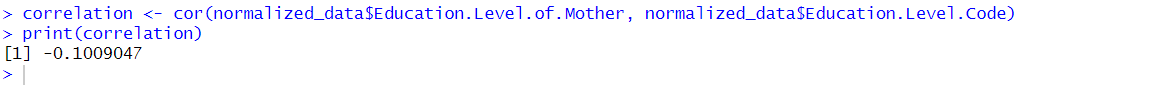
****

**Calculate the correlation between "Education.Level.of.Mother" and "Education.Level.Code":**

**Code Segment:**

correlation <- cor(normalized\_data$Education.Level.of.Mother, normalized\_data$Education.Level.Code)

print(correlation)

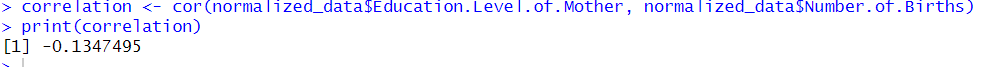


**Calculate the correlation between "Education.Level.of.Mother" and "Number.of.Births":**

**Code Segment:**

correlation <- cor(normalized\_data$Education.Level.of.Mother, normalized\_data$Number.of.Births)

print(correlation)

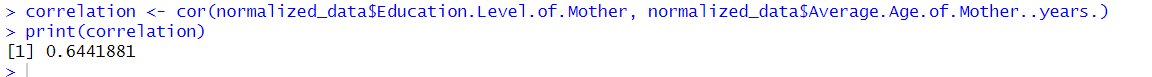
****

**Calculate the correlation between "Education.Level.of.Mother" and "Average.Age.of.Mother..years.":**

**Code Segment:**

correlation <- cor(normalized\_data$Education.Level.of.Mother, normalized\_data$Average.Age.of.Mother..years.)

print(correlation)

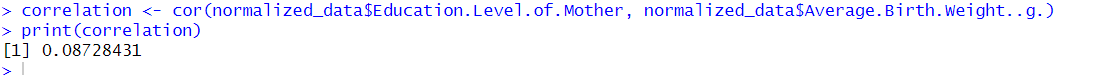
****

**Calculate the correlation between "Education.Level.of.Mother" and "Average.Birth.Weight..g.":**

**Code Segment:**

correlation <- cor(normalized\_data$Education.Level.of.Mother, normalized\_data$Average.Birth.Weight..g.)

print(correlation)

****

**Delete Column**

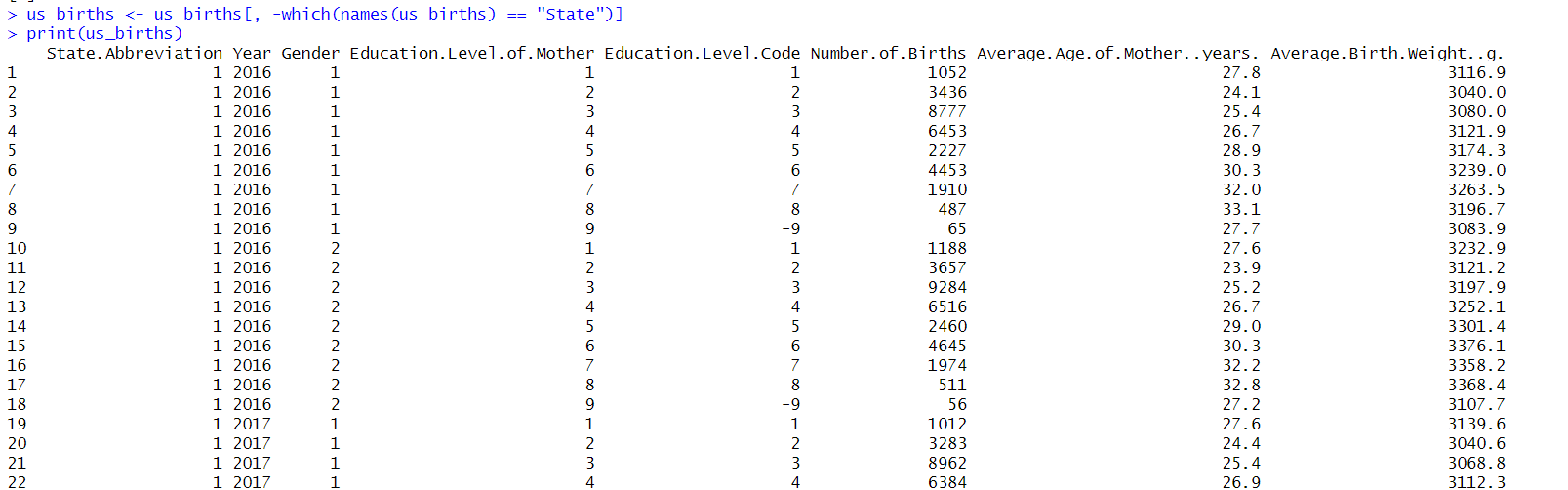
A correlation value close to 0 indicates no linear relationship between the variables, while a value close to 1 or -1 indicates a strong positive or negative linear relationship. In this dataset we get 4 column and they are state, state abbreviation, year, gender. All are close to 0 or no linear relationship.

**Delete Column (State):**

**Code Segment:**

us\_births <- us\_births[, -which(names(us\_births) == "State ")]

print(us\_births)

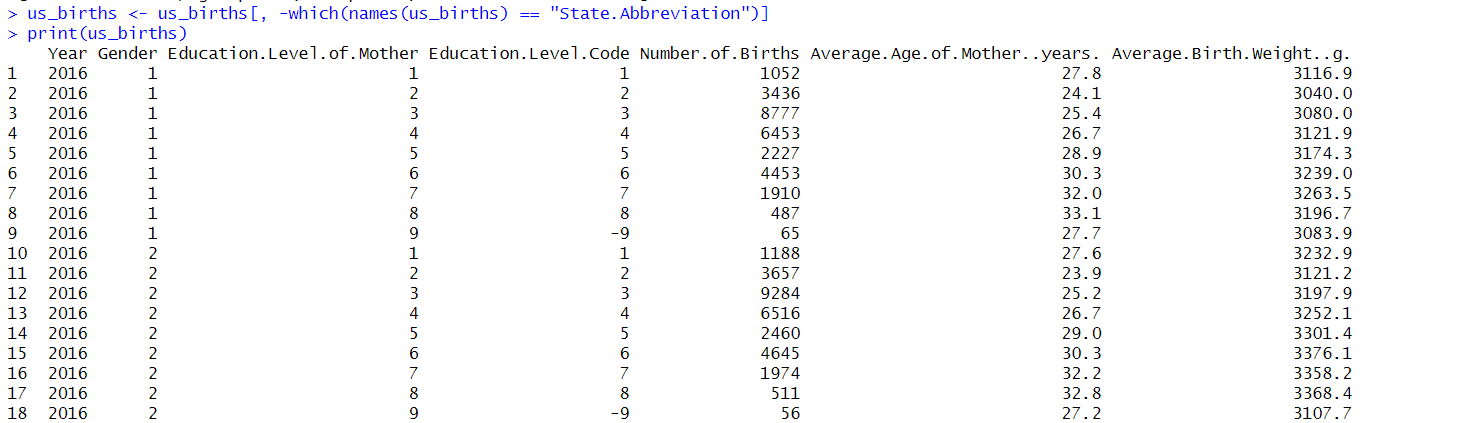
****

**Delete Column (State Abbreviation):**

**Code Segment:**

us\_births <- us\_births[, -which(names(us\_births) == "State.Abbreviation")]

print(us\_births)

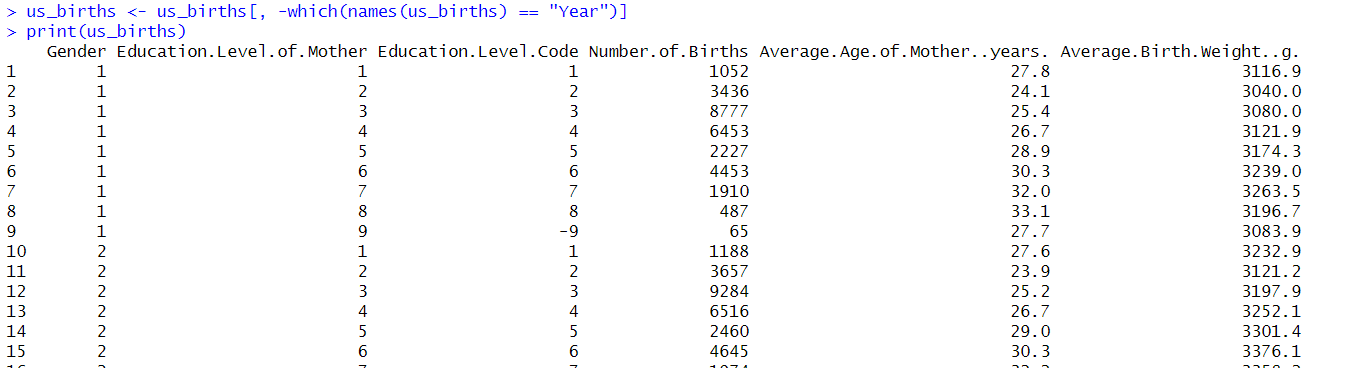
****

**Delete Column (Year):**

**Code Segment:**

us\_births <- us\_births[, -which(names(us\_births) == " Year")]

print(us\_births)

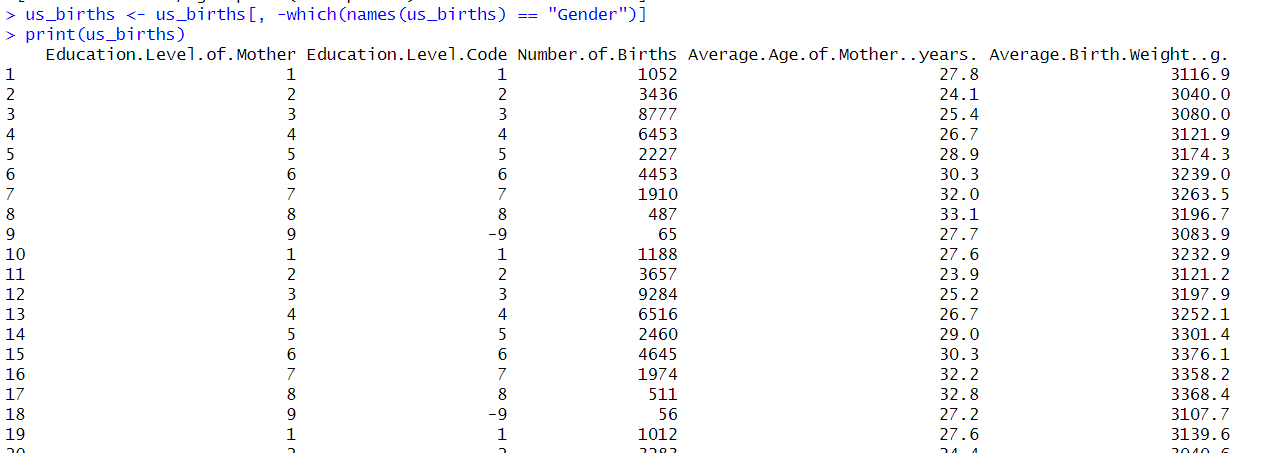
****

**Delete Column (Gender):**

**Code Segment:**

us\_births <- us\_births[, -which(names(us\_births) == " Gender ")]

print(us\_births)

****

**Plot Correlation Matrix**

A plot correlation matrix is a data visualization technique that visually represents relationships between multiple variables in a dataset. It displays correlation coefficients between pairs of variables, with color or shading indicating strength and direction. Each cell in the matrix represents the correlation between two variables, and the color or shading of the cell can be used to convey the strength and direction of the correlation.

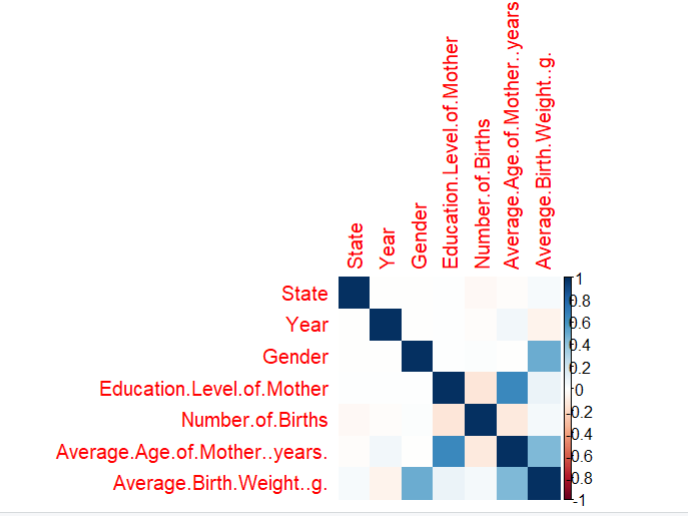
**Code Segment:**

install.packages("corrplot")

library(corrplot)

plot<-cor(normalized\_data)

corrplot(plot,method="color")



The plot is a color-coded grid where each cell represents the correlation between two variables. Blue shades represent negative correlations, while red shades represent positive correlations. The intensity of the color indicates the strength of the correlation: darker colors represent stronger correlations.

**Training & Testing**

Splitting a dataset into training and testing subsets is a crucial step in the field of data science, particularly when building and evaluating predictive models. For example: Fair Comparison, Decision Making, Validation of Results, Quality Control.

Dividing the data into training and test set.

**Code Segment:**

random <- sample(1:nrow(normalized\_data), 0.7 \* nrow(normalized\_data))

Education.Level.of.Mother\_train <- normalized\_data[random, ]

Education.Level.of.Mother\_test <- normalized\_data[-random, ]

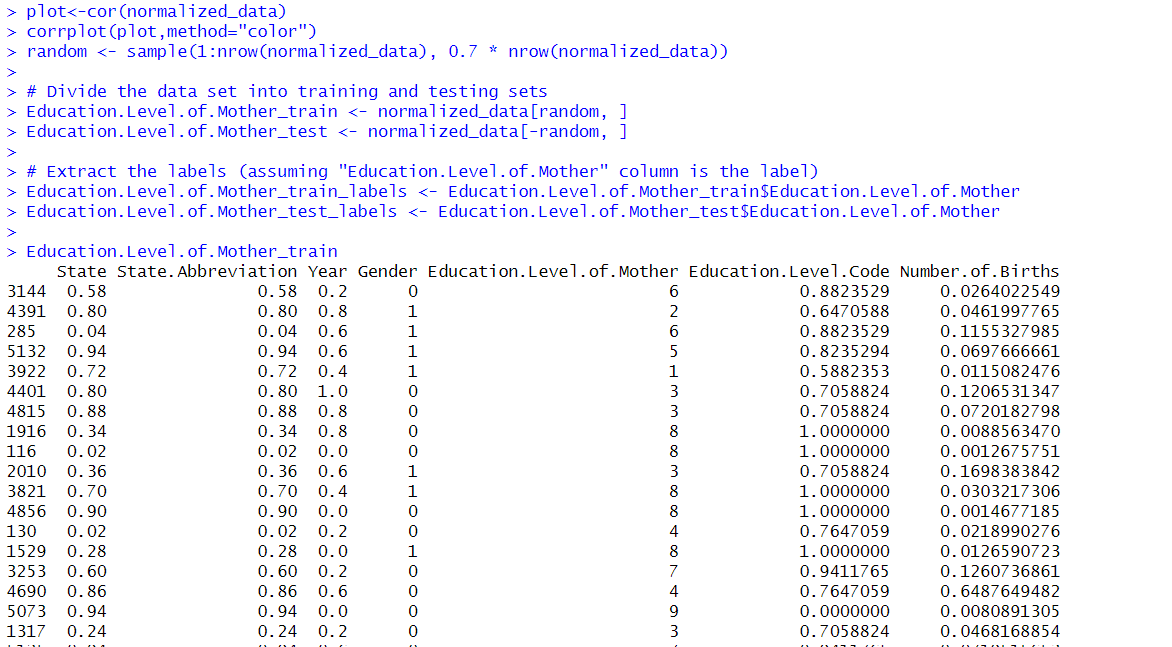
Education.Level.of.Mother\_train\_labels <- Education.Level.of.Mother\_train$Education.Level.of.Mother

Education.Level.of.Mother\_test\_labels <- Education.Level.of.Mother\_test$Education.Level.of.Mother

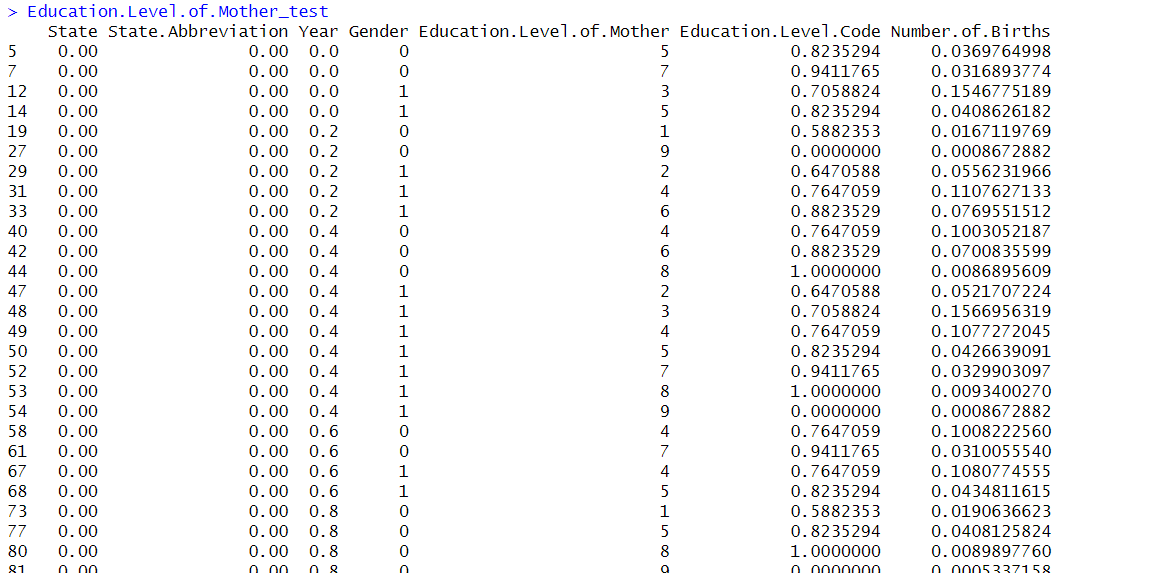
Education.Level.of.Mother\_train

Education.Level.of.Mother\_test

**For train:**



**For test:**

****

**Accuracy**

**Dividing the data into training and test set:**

In data science and machine learning, accuracy is a key metric used to measure the performance of a predictive model.

**Code Segment:**

install.packages("class")

library(class)

set.seed(123)

random <- sample(1:nrow(normalized\_data), 0.7 \* nrow(normalized\_data))

train\_data <- normalized\_data[random, ]

test\_data <- normalized\_data[-random, ]

train\_labels <- train\_data$Education.Level.of.Mother

test\_labels <- test\_data$Education.Level.of.Mother

k <- 3

knn\_model <- knn(train = train\_data[, -which(names(train\_data) == "Education.Level.of.Mother")],

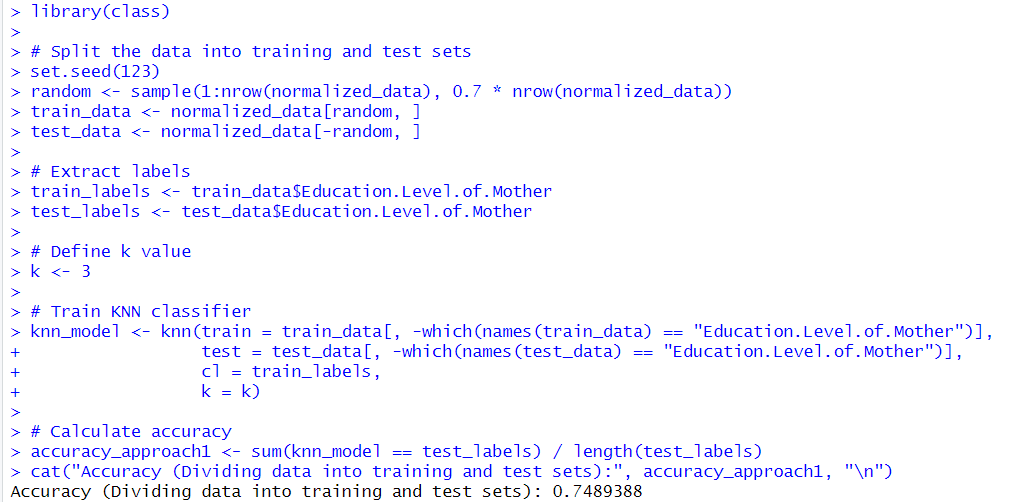
test = test\_data[, -which(names(test\_data) == "Education.Level.of.Mother")],

cl = train\_labels,

k = k)

accuracy\_approach1 <- sum(knn\_model == test\_labels) / length(test\_labels)

cat("Accuracy (Dividing data into training and test sets):", accuracy\_approach1, "\n")



**10-fold cross validation**

The 10-fold cross-validation method, which divides the dataset into 10 equal-sized subsets, is a common data science method for evaluating the effectiveness of predictive models. Its main goal is to give an accurate estimate of how well a model performs on unknown data.

**Code Segment:**

install.packages("class")

install.packages("caret")

library(class)

library(caret)

set.seed(123)

num\_folds <- 10

fold\_indices <- createFolds(normalized\_data$Education.Level.of.Mother, k = num\_folds)

accuracies <- numeric(num\_folds)

for (i in 1:num\_folds) {

test\_indices <- fold\_indices[[i]]

train\_indices <- setdiff(1:nrow(normalized\_data), test\_indices)

Education.Level.of.Mother\_train <- normalized\_data[train\_indices, ]

Education.Level.of.Mother\_test <- normalized\_data[test\_indices, ]

input\_features\_train <- Education.Level.of.Mother\_train[, c("State", "Year", "Gender","Number.of.Births", "Average.Age.of.Mother..years.", "Average.Birth.Weight..g.")]

input\_features\_test <- Education.Level.of.Mother\_test[, c("State", "Year", "Gender","Number.of.Births", "Average.Age.of.Mother..years.", "Average.Birth.Weight..g.")]

Education.Level.of.Mother\_train\_labels <- Education.Level.of.Mother\_train$Education.Level.of.Mother

Education.Level.of.Mother\_test\_labels <- Education.Level.of.Mother\_test$Education.Level.of.Mother

k <- 3 # Set the value of 'k'

predicted\_labels <- knn(train = input\_features\_train,

test = input\_features\_test,

cl = Education.Level.of.Mother\_train\_labels,

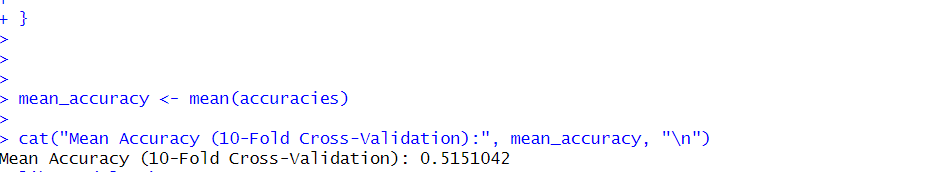
k = k)

accuracies[i] <- sum(predicted\_labels == Education.Level.of.Mother\_test\_labels) / length(Education.Level.of.Mother\_test\_labels)

}

mean\_accuracy <- mean(accuracies)

cat("Mean Accuracy (10-Fold Cross-Validation):", mean\_accuracy, "\n")



**Confusion matrix**

A confusion matrix evaluates classification model performance by comparing predicted and actual classes, revealing strengths and weaknesses, and aiding in data science.

**Code Segment:**

library(caret)

library(class)

set.seed(123)

cv <- createFolds(normalized\_data$Education.Level.of.Mother, k = 10)

for (i in 1:10) {

# Get training and testing indices for fold i

train\_indices <- cv[[i]]

test\_indices <- setdiff(1:nrow(normalized\_data), train\_indices)

input\_features\_train <- normalized\_data[train\_indices, c("State", "Year", "Gender", "Number.of.Births",

"Average.Age.of.Mother..years.", "Average.Birth.Weight..g.")]

decision\_train <- normalized\_data[train\_indices, "Education.Level.of.Mother"]

input\_features\_test <- normalized\_data[test\_indices, c("State", "Year", "Gender", "Number.of.Births",

"Average.Age.of.Mother..years.", "Average.Birth.Weight..g.")]

decision\_test <- normalized\_data[test\_indices, "Education.Level.of.Mother"]

k <- 3

predicted\_decisions <- knn(train = input\_features\_train,

test = input\_features\_test,

cl = decision\_train,

k = k)

confusion\_matrix <- table(predicted = predicted\_decisions, actual = decision\_test)

metrics <- calculate\_metrics(confusion\_matrix)

recalls[i] <- metrics$recall

precisions[i] <- metrics$precision

}

mean\_recall <- mean(recalls)

mean\_precision <- mean(precisions)

cat("Mean Recall:", mean\_recall, "\n")

cat("Mean Precision:", mean\_precision, "\n")

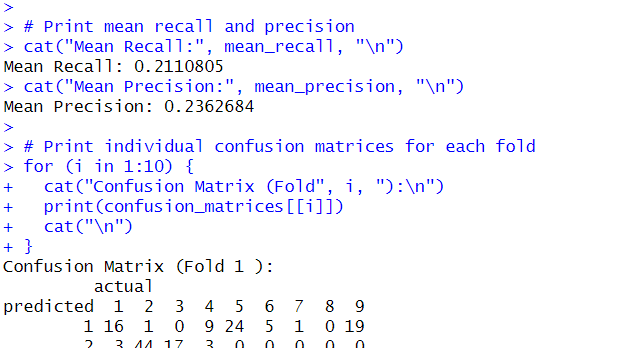
for (i in 1:10) {

cat("Confusion Matrix (Fold", i, "):\n")

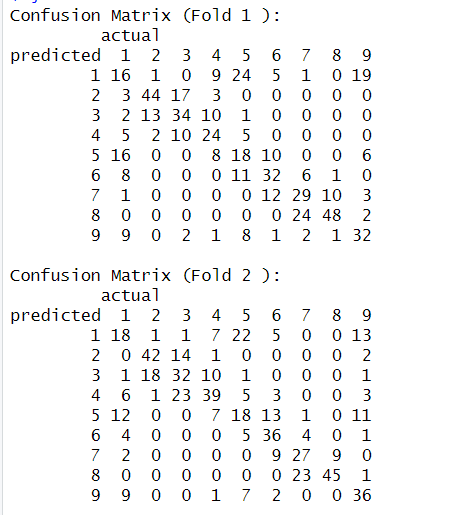
print(confusion\_matrices[[i]])

cat("\n")

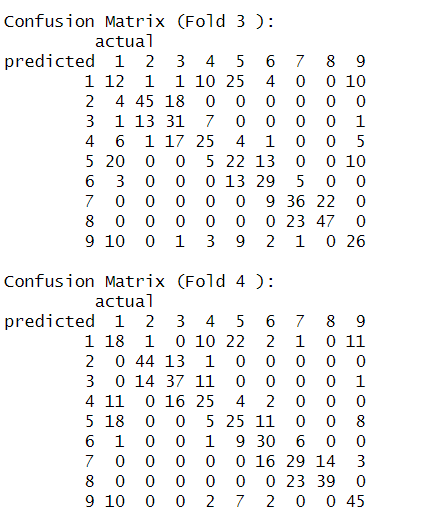
}}

****

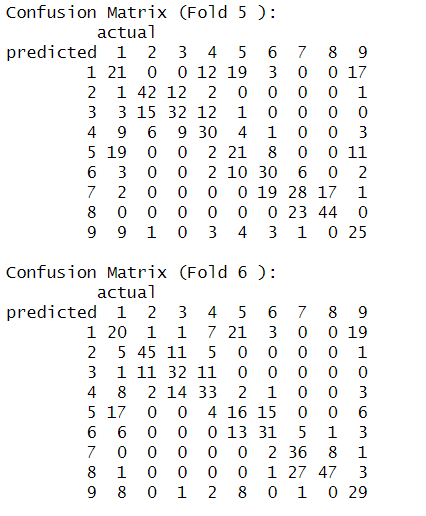
**Confusion Matrix (1,2)**

****

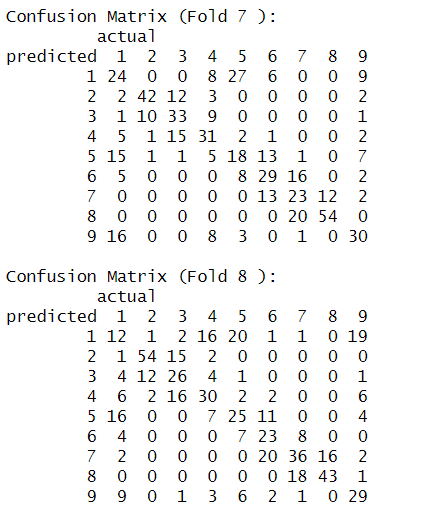
**Confusion Matrix (3,4)**

****

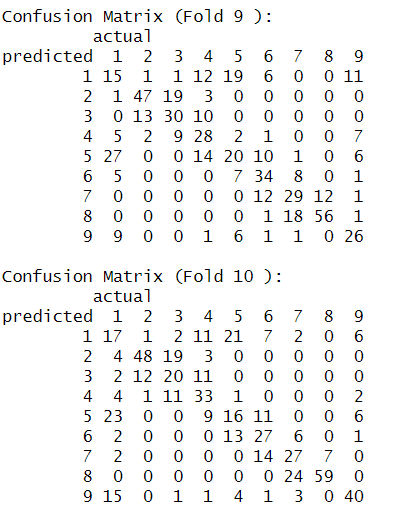
**Confusion Matrix (5,6)**

****

**Confusion Matrix (7,8)**

****

**Confusion Matrix (9,10)**

****