**6**

Implementation of Data pre-processing techniques in RStudio like, Naming and Renaming  variables, adding a new variable. Dealing with missing data. Dealing with categorical data. Data  reduction using subsetting

**Example for reference**

**Sample Data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Age** | **Gender** | **Occupation** | **Salary** |
| Alice | 30 | Female | Engineer | 60000 |
| Bob | 35 | Male | Doctor | 80000 |
| Charlie | 28 | Male | Artist |  |
| Diana |  | Female | Designer | 55000 |
| Edward | 45 | Male | Manager |  |
| Jasmine |  | Female | Writer | 45000 |

**1. Load Data from CSV (make sure to make the CSV files)**

dataframe <- read.csv("d:\\-prac6.csv")

print(dataframe)

**Naming and Renaming Variables**

**1. To Rename Variables**

names(dataframe)[names(dataframe) == "Name"] <- "FirstName"

print(dataframe)

**2. To Add a New Variable**

dataframe$emailaddress <- c('Alice@gmail.com', 'Bob@gmail.com', 'Charlie@gmail.com',  'Diana@gmail.com', 'Edward@gmail.com','Jasmine@gmail.com')

print(dataframe)

**Dealing with Missing Data**

**1. Remove Rows with Missing Values**

dataframe <- dataframe[complete.cases(dataframe), ]

print(dataframe)

**\*Load Data again from csv file to get rows with missing values back inorder to do the next statement** dataframe <- read.csv("file-prac6.csv")

**2. Fill Missing Values with Mean**

**a. To calculate the missing Age**

dataframe$Age <- ifelse(

 is.na(dataframe$Age), # Condition: If Age is NA (missing)

 ave(dataframe$Age, FUN = function(x) mean(x, na.rm = TRUE)), # Calculate mean of non missing values in Age

 dataframe$Age # Else, keep the original Age value

)

print(dataframe)

**b. To calculate the missing salary**

dataframe$Salary <- ifelse(

 is.na(dataframe$Salary), # Condition: If Salary is NA (missing)

 ave(dataframe$Salary, FUN = function(x) mean(x, na.rm = TRUE)), # Calculate mean of  non-missing values in Salary

 dataframe$Salary # Else, keep the original Salary value

)

print(dataframe)

**Dealing with Categorical Data**

**1. Convert Categorical to Numerical (using factorization)**

dataframe$Gender <- as.numeric(factor(dataframe$Gender))

print(dataframe)

**Note: Load Data again from csv file**

**Data Reduction using Subsetting**

**1. Subset Based on Condition**

subset\_data <- dataframe[dataframe$Age > 30, ]

print(subset\_data)

dataframe <- read.csv("file-prac6.csv")

print(dataframe)

**2. Random Sampling (to reduce data size for analysis)**

sample\_size <- 5 # Adjust the sample size as needed

sample\_data <- dataframe[sample(nrow(dataframe), sample\_size, replace = FALSE), ] print(sample\_data)