UNIVERSITI TUN HUSSEIN ONN MALAYSIA

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

(FSKTM)

SEMESTER II 2024/2025

DATA MINING

BIT 33603

SECTION 03

LAB ASSIGNMENT 04

**TITLE**

DATA PREPROCESSING USING R

**LECTURER'S NAME**

DR. ROZITA BINTI ABDUL JALIL

| | |
|---|---|
| NAME | TUAN KHALIDAH SYAZWANA BINTI TUAN MOHD KASMAWI |
| MATRIC NUMBER | AI220118 |
| DATE SUBMISSION | April 09, 2025 |

# LAB ACTIVITY 4

**Topic:** Data Preprocessing Using R

**Objectives:**

1. To understand basic data preprocessing techniques in R including data cleaning, encoding, and splitting datasets.
2. To apply missing value imputation and data normalization methods to prepare data for analysis.

**Duration:** 2 hours

---

**Assessment Question:**

1. Run the provided code in R (Activity 1-4) and understanding the data preprocessing.
2. Submit the visualizations as image/data snapshots for each activity (before and after) along with a brief explanation of the insights gained.

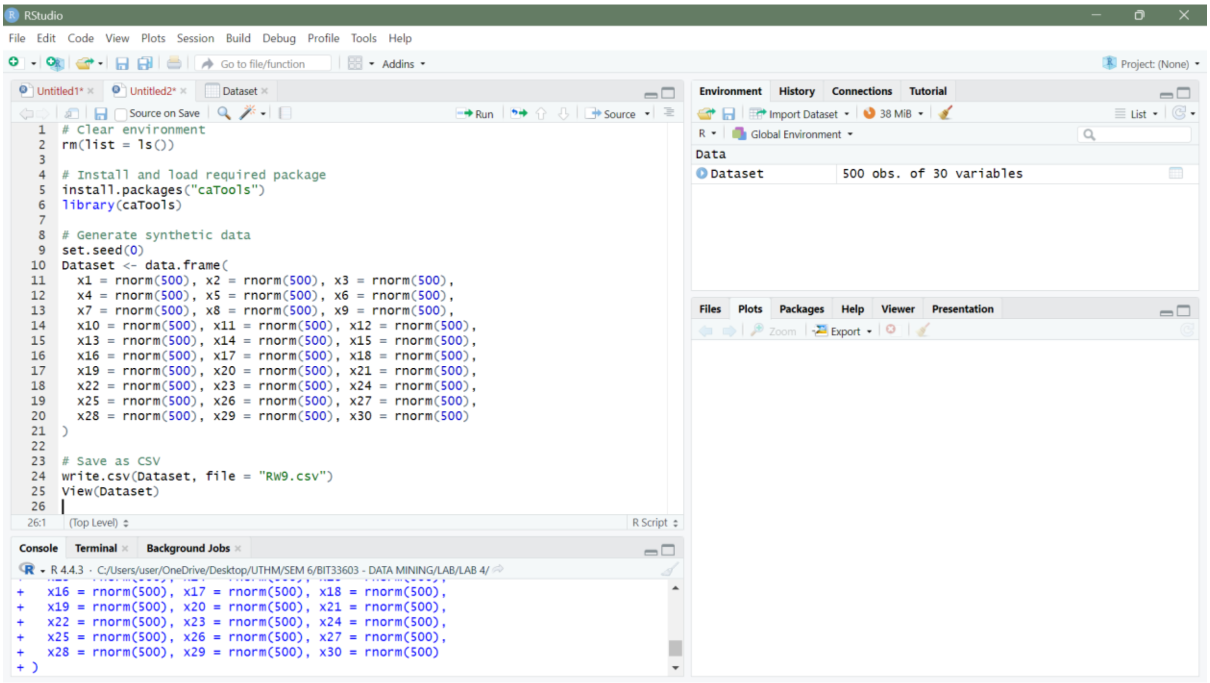
---

**Submission Guidelines:**

1. Submit your solution/answer as a report or document in a single file (.pdf or .docx format).
2. Include a cover page that contains your name, matrix number, and lab name.
3. On the following page, insert screenshots of each activity.
4. Submit your lab exercise through AUTHOR.

## Activity 1: Creating a Synthetic Dataset

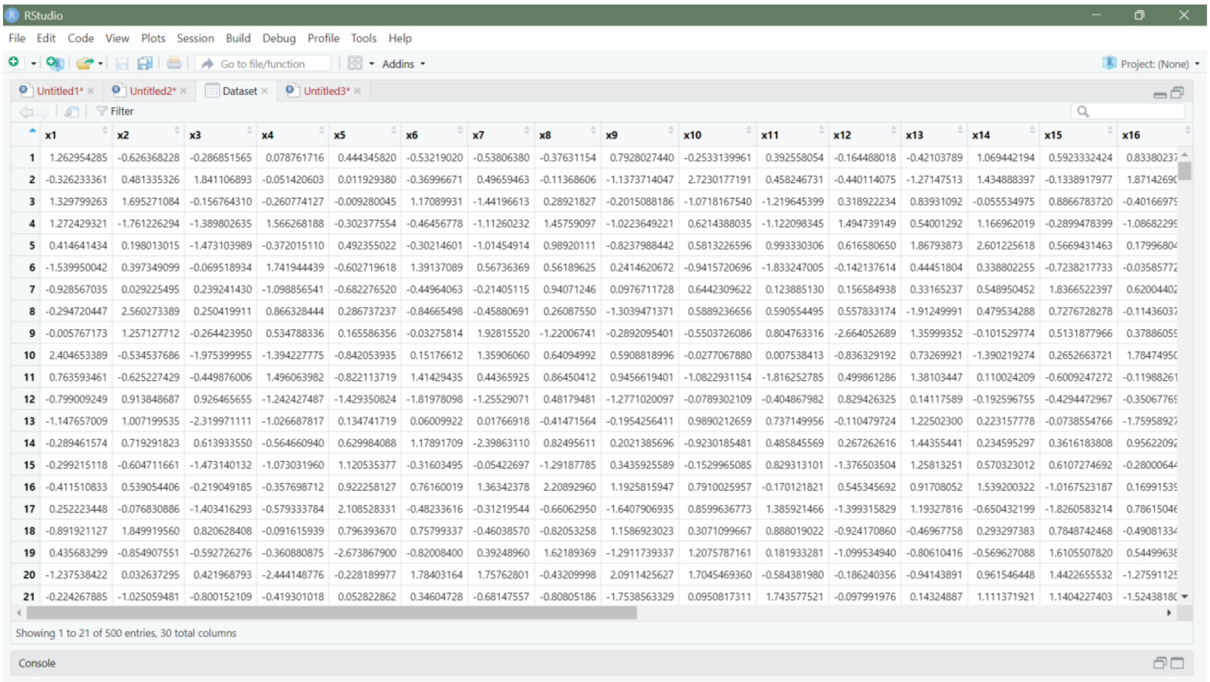
In this activity, you will generate a synthetic dataset using random value  rnorm() and save it as a CSV file. This exercise helps you understand how data can be simulated for practice in data mining techniques.



```r
# Clear environment
rm(list = ls())

# Install and load required package
install.packages("caTools")
library(caTools)

# Generate synthetic data
set.seed(0)
Dataset <- data.frame(
  x1 = rnorm(500), x2 = rnorm(500), x3 = rnorm(500),
  x4 = rnorm(500), x5 = rnorm(500), x6 = rnorm(500),
  x7 = rnorm(500), x8 = rnorm(500), x9 = rnorm(500),
  x10 = rnorm(500), x11 = rnorm(500), x12 = rnorm(500),
  x13 = rnorm(500), x14 = rnorm(500), x15 = rnorm(500),
  x16 = rnorm(500), x17 = rnorm(500), x18 = rnorm(500),
  x19 = rnorm(500), x20 = rnorm(500), x21 = rnorm(500),
  x22 = rnorm(500), x23 = rnorm(500), x24 = rnorm(500),
  x25 = rnorm(500), x26 = rnorm(500), x27 = rnorm(500),
  x28 = rnorm(500), x29 = rnorm(500), x30 = rnorm(500)
)

# Save as CSV
write.csv(Dataset, file = "RW9.csv")
View(Dataset)
```

Environment — Global Environment — Data: Dataset 500 obs. of 30 variables

```
R 4.4.3 · C:/Users/user/OneDrive/Desktop/UTHM/SEM 6/BIT33603 - DATA MINING/LAB/LAB 4/
+   x16 = rnorm(500), x17 = rnorm(500), x18 = rnorm(500),
+   x19 = rnorm(500), x20 = rnorm(500), x21 = rnorm(500),
+   x22 = rnorm(500), x23 = rnorm(500), x24 = rnorm(500),
+   x25 = rnorm(500), x26 = rnorm(500), x27 = rnorm(500),
+   x28 = rnorm(500), x29 = rnorm(500), x30 = rnorm(500)
+ )
```

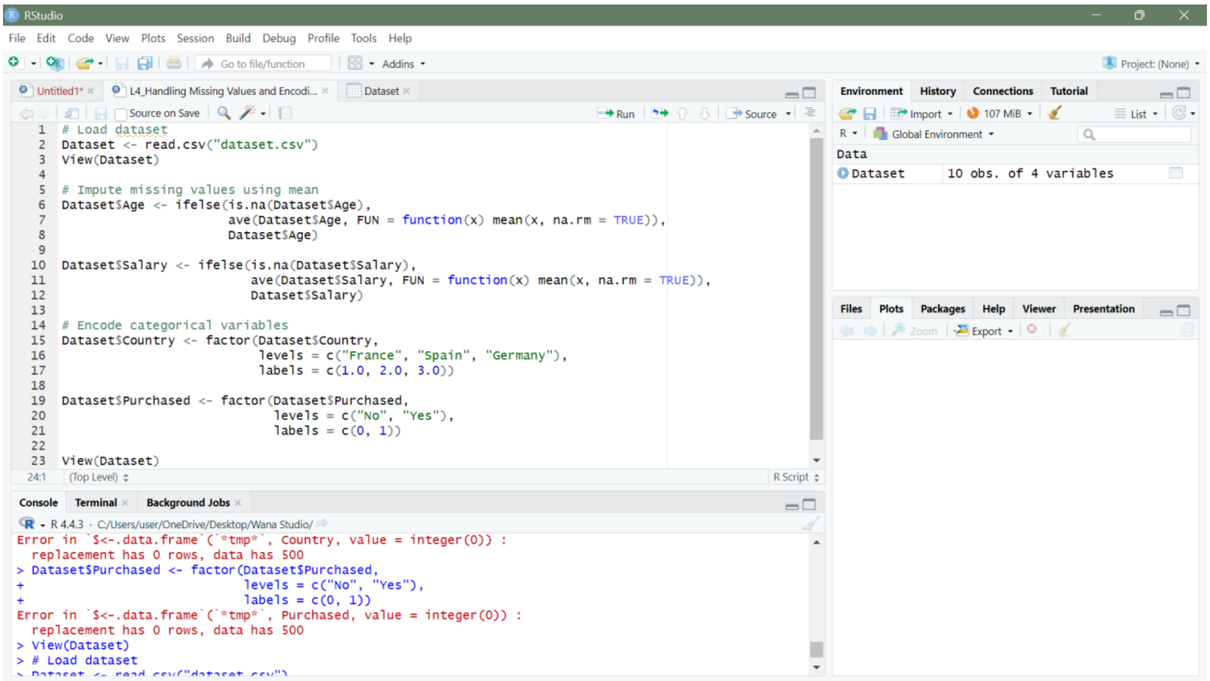


| | x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8 | x9 | x10 | x11 | x12 | x13 | x14 | x15 | x16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.262954285 | -0.626368228 | -0.286851565 | 0.078761716 | 0.444345820 | -0.53219020 | -0.53806380 | -0.37631154 | 0.7928027440 | -0.2533139961 | 0.392558054 | -0.164488018 | -0.42103789 | 1.069442194 | 0.5923332424 | 0.83380237 |
| 2 | -0.326233361 | 0.481335326 | 1.841106893 | -0.051420603 | 0.011929380 | -0.36996671 | 0.49659463 | -0.11368606 | -1.1373714047 | 2.7230177191 | 0.458246731 | -0.440114075 | -1.27147513 | 1.434888397 | -0.1338917977 | 1.8714269 |
| 3 | 1.329799263 | 1.695271084 | -0.156764310 | -0.260774127 | -0.009280045 | 1.17089931 | -1.44196613 | 0.28921827 | -0.2015088186 | -1.0718167540 | -1.219645399 | 0.318922234 | 0.83931092 | -0.055534975 | 0.8866783720 | -0.4016697 |
| 4 | 1.272429321 | -1.761226294 | -1.389802635 | 1.566268188 | -0.302377554 | -0.46456778 | -1.11260232 | 1.45759097 | -1.0223649221 | 0.6214388035 | -1.122098345 | 1.494739149 | 0.54001292 | 1.166962019 | -0.2899478399 | -1.0868229 |
| 5 | 0.414641434 | 0.198013015 | -1.473103989 | -0.372015110 | 0.492355022 | -0.30214601 | -1.01454914 | 0.98920111 | -0.8237988442 | 0.5813226596 | 0.993330306 | 0.616580650 | 1.86793873 | 2.601225618 | 0.5669431463 | 0.1799680 |
| 6 | -1.539950042 | 0.397349099 | -0.069518934 | 1.741944439 | -0.602719618 | 1.39137089 | 0.56736369 | 0.56189625 | 0.2414620672 | -0.9415720696 | -1.833247005 | -0.142137614 | 0.44451804 | 0.338802255 | -0.7238217733 | -0.0358577 |
| 7 | -0.928567035 | 0.029225495 | 0.239241430 | -1.098856541 | -0.682276520 | -0.44964063 | -0.21405115 | 0.94071246 | 0.0976711728 | 0.6442309622 | 0.123885130 | 0.156584938 | 0.33165237 | 0.548950452 | 1.8366522397 | 0.6200440 |
| 8 | -0.294720447 | 2.560273389 | 0.250419911 | 0.866328444 | 0.286737237 | -0.84665498 | -0.45880691 | 0.26087550 | -1.3039471371 | 0.5889236656 | 0.590554495 | 0.557833174 | -1.91249991 | 0.479534288 | 0.7276728278 | -0.1143603 |
| 9 | -0.005767173 | 1.257127712 | -0.264423950 | 0.534788336 | 0.165586356 | -0.03275814 | 1.92815520 | -1.22006741 | -0.2892095401 | -0.5503726086 | 0.804763316 | -2.664052689 | 1.35999352 | -0.101529774 | 0.5131877966 | 0.3788605 |
| 10 | 2.404653389 | -0.534537686 | -1.975399955 | -1.394227775 | -0.842053935 | 0.15176612 | 1.35906060 | 0.64094992 | 0.5908818996 | -0.0277067880 | 0.007538413 | -0.836329192 | 0.73269921 | -1.390219274 | 0.2652663721 | 1.7847495 |
| 11 | 0.763593461 | -0.625227429 | -0.449876006 | 1.496063982 | -0.822113719 | 1.41429435 | 0.44365925 | 0.86450412 | 0.9456619401 | -1.0822931154 | -1.816252785 | 0.499861286 | 1.38103447 | 0.110024209 | -0.6009247272 | -0.1198826 |
| 12 | -0.799009249 | 0.913848687 | 0.926465655 | -1.242427487 | -1.429350824 | -1.81978098 | -1.25529071 | 0.48179481 | -1.2771020097 | -0.0789302109 | -0.404867982 | 0.829426325 | 0.14117589 | -0.192596755 | -0.4294472967 | -0.3506776 |
| 13 | -1.147657009 | 1.007199535 | -2.319971111 | -1.026687817 | 0.134741719 | 0.06009922 | 0.01766918 | -0.41471564 | -0.1954256411 | 0.9890212659 | 0.737149956 | -0.110479724 | 1.22502300 | 0.223157778 | -0.0738554766 | -1.7595892 |
| 14 | -0.289461574 | 0.719291823 | 0.613933550 | -0.564660940 | 0.629984088 | 1.17891709 | -2.39863110 | 0.82495611 | 0.2021385696 | -0.9230185481 | 0.485845569 | 0.267262616 | 1.44355441 | 0.234595297 | 0.3616183808 | 0.9562209 |
| 15 | -0.299215118 | -0.604711661 | -1.473140132 | -1.073031960 | 1.120535377 | -0.31603495 | -0.05422697 | -1.29187785 | 0.3435925589 | -0.1529965085 | 0.829313101 | -1.376503504 | 1.25813251 | 0.570323012 | 0.6107274692 | -0.2800064 |
| 16 | -0.411510833 | 0.539054406 | -0.219049185 | -0.357698712 | 0.922258127 | 0.76160019 | 1.36342378 | 2.20892960 | 1.1925815947 | 0.7910025957 | -0.170121821 | 0.545345692 | 0.91708052 | 1.539200322 | -1.0167523187 | 0.1699153 |
| 17 | 0.252223448 | -0.076830886 | -1.403416293 | -0.579333784 | 2.108528331 | -0.48233616 | -0.31219544 | -0.66062950 | -1.6407906935 | 0.8599636773 | 1.385921466 | -1.399315829 | 1.19327816 | -0.650432199 | -1.8260583214 | 0.7861504 |
| 18 | -0.891921127 | 1.849919560 | 0.820628408 | -0.091615939 | 0.796393670 | 0.75799337 | -0.46038570 | -0.82053258 | 1.1586923023 | 0.3071099667 | 0.888019022 | -0.924170860 | -0.46967758 | 0.293297383 | 0.7848742468 | -0.4908133 |
| 19 | 0.435683299 | -0.854907551 | -0.592726276 | -0.360880875 | -2.673867900 | -0.82008400 | 0.39248960 | 1.62189369 | -1.2911739337 | 1.2075787161 | 0.181933281 | -1.099534940 | -0.80610416 | -0.569627088 | 1.6105507820 | 0.5449963 |
| 20 | -1.237538422 | 0.032637295 | 0.421968793 | -2.444148776 | -0.228189977 | 1.78403164 | 1.75762801 | -0.43209998 | 2.0911425627 | 1.7045469360 | -0.584381980 | -0.186240356 | -0.94143891 | 0.961546448 | 1.4422655532 | -1.2759112 |
| 21 | -0.224267885 | -1.025059481 | -0.800152109 | -0.419301018 | 0.052822862 | 0.34604728 | -0.68147557 | -0.80805186 | -1.7538563329 | 0.0950817311 | 1.743577521 | -0.097991976 | 0.14324887 | 1.111371921 | 1.1404227403 | -1.5243818 |

Showing 1 to 21 of 500 entries, 30 total columns

## Activity 2: Handling Missing Values and Encoding Categorical Data

In this activity, you will learn to load an existing dataset, detect and handle missing values using mean imputation, and encode categorical variables into numeric form using factor the factor() function.



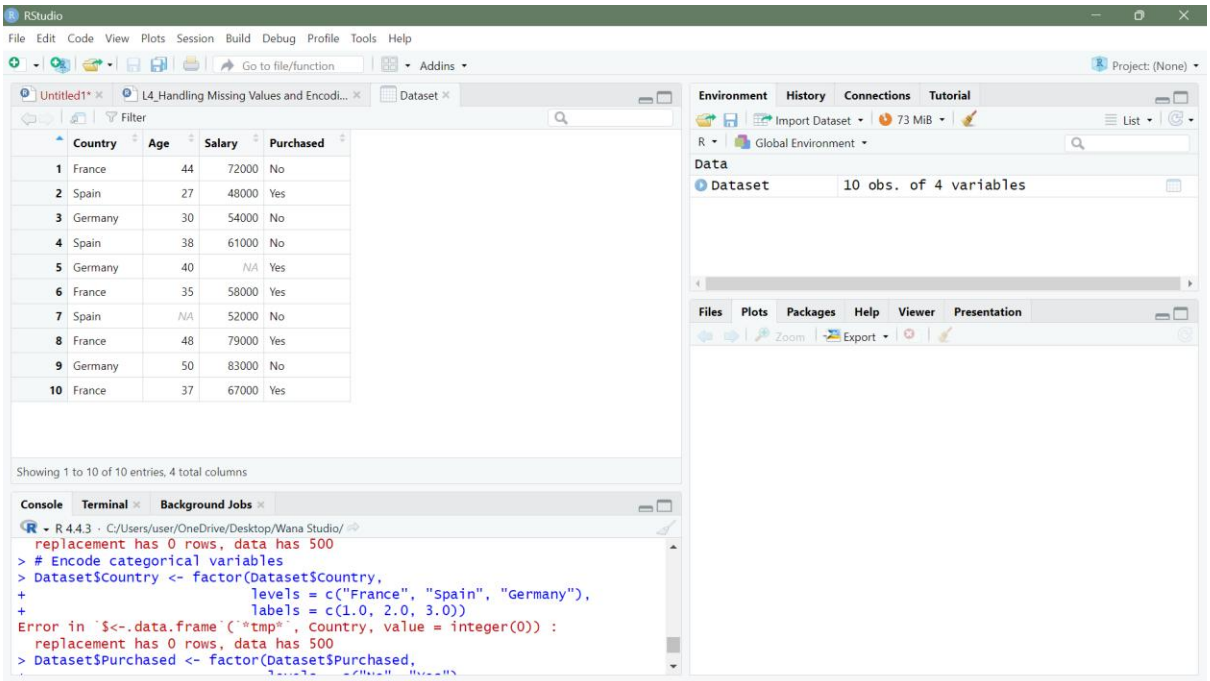
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Untitled1* L4_Handling Missing Values and Encodi... Dataset
Source on Save
Run
Source
Environment History Connections Tutorial
Import 107 MiB
List
Global Environment
Files Plots Packages Help Viewer Presentation
Zoom Export
Console Terminal Background Jobs
R Script
(Top Level)
24:1

```
# Load dataset
Dataset <- read.csv("dataset.csv")
View(Dataset)

# Impute missing values using mean
Dataset$Age <- ifelse(is.na(Dataset$Age),
                      ave(Dataset$Age, FUN = function(x) mean(x, na.rm = TRUE)),
                      Dataset$Age)

Dataset$Salary <- ifelse(is.na(Dataset$Salary),
                         ave(Dataset$Salary, FUN = function(x) mean(x, na.rm = TRUE)),
                         Dataset$Salary)

# Encode categorical variables
Dataset$Country <- factor(Dataset$Country,
                          levels = c("France", "Spain", "Germany"),
                          labels = c(1.0, 2.0, 3.0))

Dataset$Purchased <- factor(Dataset$Purchased,
                            levels = c("No", "Yes"),
                            labels = c(0, 1))

View(Dataset)
```

Data
Dataset 10 obs. of 4 variables

R 4.4.3 · C:/Users/user/OneDrive/Desktop/Wana Studio/

```
Error in `$<-.data.frame`(`*tmp*`, Country, value = integer(0)) :
  replacement has 0 rows, data has 500
> Dataset$Purchased <- factor(Dataset$Purchased,
+                             levels = c("No", "Yes"),
+                             labels = c(0, 1))
Error in `$<-.data.frame`(`*tmp*`, Purchased, value = integer(0)) :
  replacement has 0 rows, data has 500
> View(Dataset)
> # Load dataset
> Dataset <- read.csv("dataset.csv")
```



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1* L4_Handling Missing Values and Encodi... Dataset
Filter
Environment History Connections Tutorial
Import Dataset 73 MiB
List
Global Environment
Files Plots Packages Help Viewer Presentation
Zoom Export
Console Terminal Background Jobs

| | Country | Age | Salary | Purchased |
|---|---|---|---|---|
| 1 | France | 44 | 72000 | No |
| 2 | Spain | 27 | 48000 | Yes |
| 3 | Germany | 30 | 54000 | No |
| 4 | Spain | 38 | 61000 | No |
| 5 | Germany | 40 | NA | Yes |
| 6 | France | 35 | 58000 | Yes |
| 7 | Spain | NA | 52000 | No |
| 8 | France | 48 | 79000 | Yes |
| 9 | Germany | 50 | 83000 | No |
| 10 | France | 37 | 67000 | Yes |

Showing 1 to 10 of 10 entries, 4 total columns

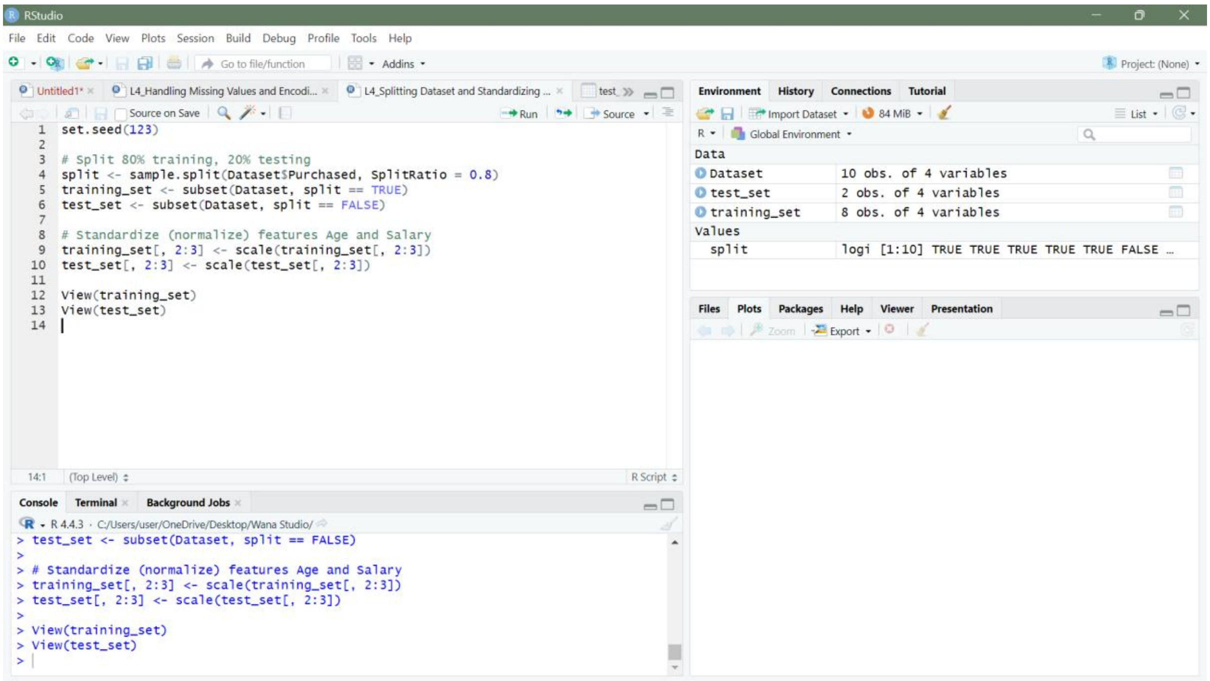
Data
Dataset 10 obs. of 4 variables

R 4.4.3 · C:/Users/user/OneDrive/Desktop/Wana Studio/

```
  replacement has 0 rows, data has 500
> # Encode categorical variables
> Dataset$Country <- factor(Dataset$Country,
+                           levels = c("France", "Spain", "Germany"),
+                           labels = c(1.0, 2.0, 3.0))
Error in `$<-.data.frame`(`*tmp*`, Country, value = integer(0)) :
  replacement has 0 rows, data has 500
> Dataset$Purchased <- factor(Dataset$Purchased,
```

## Activity 3: Splitting Dataset and Standardizing Features

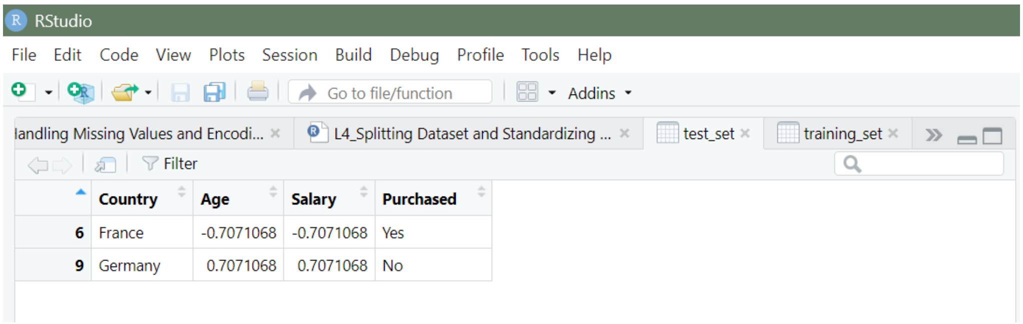
You will split the dataset into training and testing subsets (80/20 split) and apply feature scaling using scale().


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
L4_Handling Missing Values and Encodi...
L4_Splitting Dataset and Standardizing ...
test_
Source on Save
Run
Source
14:1
(Top Level)
R Script
Environment History Connections Tutorial
Import Dataset
84 MiB
List
R
Global Environment
Files Plots Packages Help Viewer Presentation
Zoom
Export
Console Terminal Background Jobs
R 4.4.3 · C:/Users/user/OneDrive/Desktop/Wana Studio/

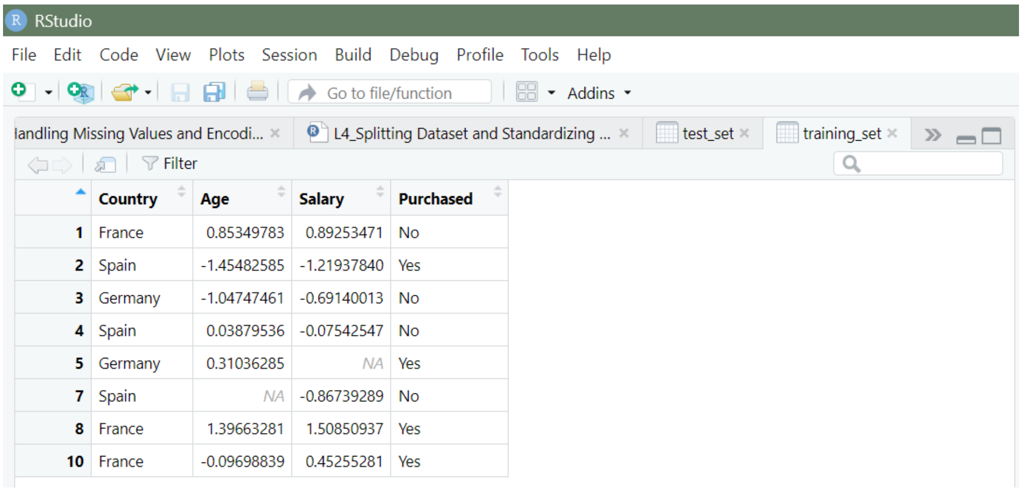
```r
set.seed(123)

# Split 80% training, 20% testing
split <- sample.split(Dataset$Purchased, SplitRatio = 0.8)
training_set <- subset(Dataset, split == TRUE)
test_set <- subset(Dataset, split == FALSE)

# Standardize (normalize) features Age and Salary
training_set[, 2:3] <- scale(training_set[, 2:3])
test_set[, 2:3] <- scale(test_set[, 2:3])

View(training_set)
View(test_set)
```

```
Data
Dataset        10 obs. of 4 variables
test_set       2 obs. of 4 variables
training_set   8 obs. of 4 variables
Values
split          logi [1:10] TRUE TRUE TRUE TRUE TRUE FALSE …
```

```
> test_set <- subset(Dataset, split == FALSE)
>
> # Standardize (normalize) features Age and Salary
> training_set[, 2:3] <- scale(training_set[, 2:3])
> test_set[, 2:3] <- scale(test_set[, 2:3])
>
> View(training_set)
> View(test_set)
> 
```


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
landling Missing Values and Encodi...
L4_Splitting Dataset and Standardizing ...
test_set
training_set
Filter

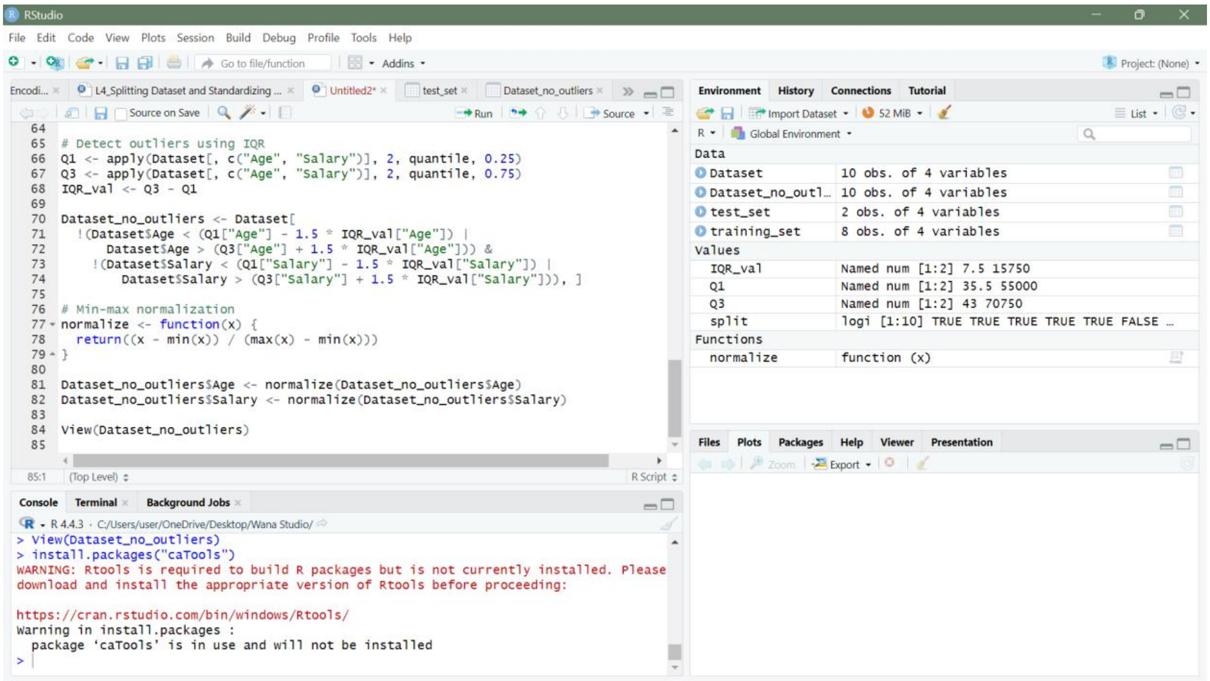
| | Country | Age | Salary | Purchased |
|---|---|---|---|---|
| 6 | France | -0.7071068 | -0.7071068 | Yes |
| 9 | Germany | 0.7071068 | 0.7071068 | No |


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
landling Missing Values and Encodi...
L4_Splitting Dataset and Standardizing ...
test_set
training_set
Filter

| | Country | Age | Salary | Purchased |
|---|---|---|---|---|
| 1 | France | 0.85349783 | 0.89253471 | No |
| 2 | Spain | -1.45482585 | -1.21937840 | Yes |
| 3 | Germany | -1.04747461 | -0.69140013 | No |
| 4 | Spain | 0.03879536 | -0.07542547 | No |
| 5 | Germany | 0.31036285 | NA | Yes |
| 7 | Spain | NA | -0.86739289 | No |
| 8 | France | 1.39663281 | 1.50850937 | Yes |
| 10 | France | -0.09698839 | 0.45255281 | Yes |

## Activity 4: Handling outlier and Normalization

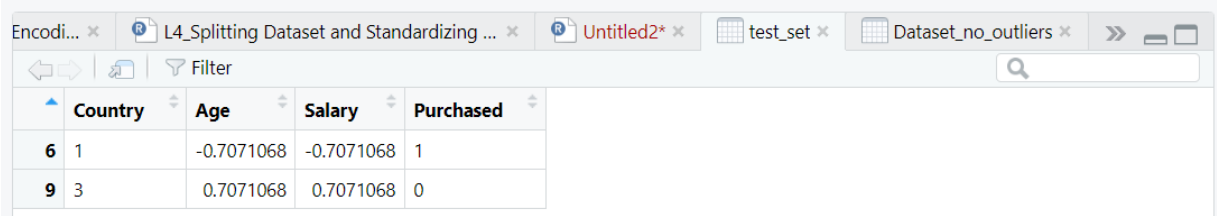
Detect and remove outliers using the IQR method. Then, apply min-max normalization.



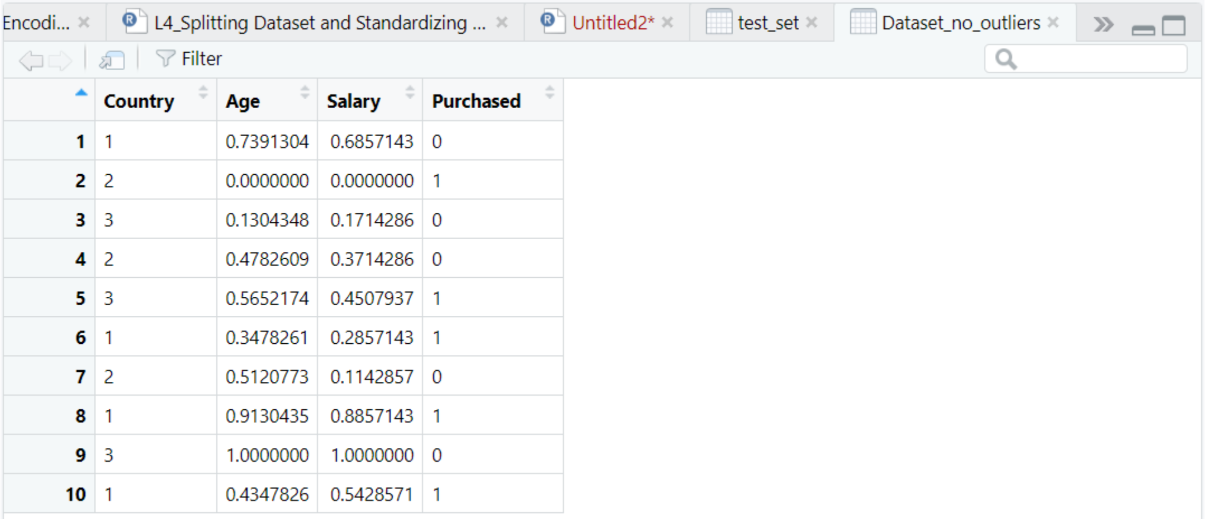
```r

# Detect outliers using IQR
Q1 <- apply(Dataset[, c("Age", "Salary")], 2, quantile, 0.25)
Q3 <- apply(Dataset[, c("Age", "Salary")], 2, quantile, 0.75)
IQR_val <- Q3 - Q1

Dataset_no_outliers <- Dataset[
  !(Dataset$Age < (Q1["Age"] - 1.5 * IQR_val["Age"]) |
      Dataset$Age > (Q3["Age"] + 1.5 * IQR_val["Age"])) &
    !(Dataset$Salary < (Q1["Salary"] - 1.5 * IQR_val["Salary"]) |
        Dataset$Salary > (Q3["Salary"] + 1.5 * IQR_val["Salary"])), ]

# Min-max normalization
normalize <- function(x) {
  return((x - min(x)) / (max(x) - min(x)))
}

Dataset_no_outliers$Age <- normalize(Dataset_no_outliers$Age)
Dataset_no_outliers$Salary <- normalize(Dataset_no_outliers$Salary)

View(Dataset_no_outliers)

```

```
Data
Dataset              10 obs. of 4 variables
Dataset_no_outl…     10 obs. of 4 variables
test_set             2 obs. of 4 variables
training_set         8 obs. of 4 variables
Values
  IQR_val            Named num [1:2] 7.5 15750
  Q1                 Named num [1:2] 35.5 55000
  Q3                 Named num [1:2] 43 70750
  split              logi [1:10] TRUE TRUE TRUE TRUE TRUE FALSE …
Functions
  normalize          function (x)
```

```
R 4.4.3 · C:/Users/user/OneDrive/Desktop/Wana Studio/
> View(Dataset_no_outliers)
> install.packages("caTools")
WARNING: Rtools is required to build R packages but is not currently installed. Please
download and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Warning in install.packages :
  package ‘caTools’ is in use and will not be installed
>
```



| | Country | Age | Salary | Purchased |
|---|---|---|---|---|
| 6 | 1 | -0.7071068 | -0.7071068 | 1 |
| 9 | 3 | 0.7071068 | 0.7071068 | 0 |



| | Country | Age | Salary | Purchased |
|---|---|---|---|---|
| 1 | 1 | 0.7391304 | 0.6857143 | 0 |
| 2 | 2 | 0.0000000 | 0.0000000 | 1 |
| 3 | 3 | 0.1304348 | 0.1714286 | 0 |
| 4 | 2 | 0.4782609 | 0.3714286 | 0 |
| 5 | 3 | 0.5652174 | 0.4507937 | 1 |
| 6 | 1 | 0.3478261 | 0.2857143 | 1 |
| 7 | 2 | 0.5120773 | 0.1142857 | 0 |
| 8 | 1 | 0.9130435 | 0.8857143 | 1 |
| 9 | 3 | 1.0000000 | 1.0000000 | 0 |
| 10 | 1 | 0.4347826 | 0.5428571 | 1 |