

#Data Preprocessing

```
#install.packages("readr")
```

```
library(readr)
```

```
#Importing the dataset
```

```
df = read.csv('Dataset1.csv')
```

```
View(df)
```

```
#Handling the missing data
```

```
#NA- no value is available
```

```
#Replace the missing data with the average of the feature in which the data
```

```
#is missing:
```

```
df$Age = ifelse(is.na(df$Age),  
               ave(df$Age, FUN = function (x)mean(x, na.rm = TRUE)),  
               df$Age)
```

```
View(df)
```

```
df$Salary = ifelse(is.na(df$Salary),  
                  ave(df$Salary, FUN = function (x)mean(x, na.rm = TRUE)),  
                  df$Salary)
```

```
View(df)
```

```
#Encoding categorical data
```

```
#Encoding refers to transforming text data into numeric data
```

```
#To transform a categorical variable into numeric, use the factor() function.
```

```
df$Country = factor(df$Country,
```

```
levels = c('France','Spain','Germany'),  
labels = c(1.0, 2.0 , 3.0 ))
```

```
View(df)
```

```
df$Purchased = factor(df$Purchased,levels = c('No', 'Yes'),labels = c(0, 1))
```

```
View(df)
```

```
df$Purchased[is.na(df$Purchased)] <- 0
```

```
View(df)
```

```
as.factor(df$Purchased)
```

```
View(df)
```

```
#Splitting the data set into the training and test set
```

```
library(caTools)# required library for data splitting
```

```
split = sample.split(df$Purchased, SplitRatio = 0.8)# returns true if observation goes to the Training  
set and false if observation goes to the test set.
```

```
#Creating the training set and test set separately
```

```
training_set = subset(df, split == TRUE)
```

```
test_set = subset(df, split == FALSE)
```

```
training_set
```

```
test_set
```

```
training_set[, 2:3] = scale(training_set[, 2:3])
```

```
test_set[, 2:3]
```

```
test_set[, 2:3] = scale(test_set[, 2:3])
```

```
training_set
```

```
test_set
```

#Scale() is a built-in R function that centers and/or scales the columns of a numeric matrix by default.