This is a template for data preprocessing

========== Importing the dataset ========

```
dataset <-  read.csv('Data.csv', header = TRUE, sep = ',')
print (dataset)
```

```
##    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
```

======== Taking care of the missing data ========

```
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)
print (dataset)
```

```
##    Country      Age   Salary Purchased
## 1   France 44.00000 72000.00        No
## 2    Spain 27.00000 48000.00       Yes
## 3  Germany 30.00000 54000.00        No
## 4    Spain 38.00000 61000.00        No
## 5  Germany 40.00000 63777.78       Yes
## 6   France 35.00000 58000.00       Yes
## 7    Spain 38.77778 52000.00        No
## 8   France 48.00000 79000.00       Yes
## 9  Germany 50.00000 83000.00        No
## 10  France 37.00000 67000.00       Yes
```

========== Taking care of the categorical data ======

```
dataset$Country = factor(dataset$Country, levels = c('France',
                  'Germany', 'Spain'), labels =c(1, 2, 3))

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

```
##    Country      Age   Salary Purchased
## 1        1 44.00000 72000.00         0
```

```
## 2        3 27.00000 48000.00         1
## 3        2 30.00000 54000.00         0
## 4        3 38.00000 61000.00         0
## 5        2 40.00000 63777.78         1
## 6        1 35.00000 58000.00         1
## 7        3 38.77778 52000.00         0
## 8        1 48.00000 79000.00         1
## 9        2 50.00000 83000.00         0
## 10       1 37.00000 67000.00         1
```

========== Splitting the data into training and testing sets

```
library(caTools)
set.seed(123)
split <-  sample.split(dataset$Purchased, SplitRatio = 0.8)
training_set <-  subset(dataset, split == TRUE)
test_set <-  subset(dataset, split == FALSE)
```

========= Feature scaling =========

```
training_set[, 2:3] <- scale(training_set[, 2:3]) #factors are
                                #not numerical type
test_set[, 2:3] <- scale(test_set[, 2:3])
print('Training_set: ')
```

```
## [1] "Training_set: "
```

```
print(training_set)
```

```
##    Country         Age     Salary Purchased
## 1        1  0.90101716  0.9392746         0
## 2        3 -1.58847494 -1.3371160         1
## 3        2 -1.14915281 -0.7680183         0
## 4        3  0.02237289 -0.1040711         0
## 5        2  0.31525431  0.1594000         1
## 7        3  0.13627122 -0.9577176         0
## 8        1  1.48678000  1.6032218         1
## 10       1 -0.12406783  0.4650265         1
```

```
print('Test_set: ')
```

```
## [1] "Test_set: "
```

```
print(test_set)
```

```
##   Country        Age     Salary Purchased
## 6       1 -0.7071068 -0.7071068         1
## 9       2  0.7071068  0.7071068         0
```