Encoding_R

```
dataset = read.csv(file='Data.csv')
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
```

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

```
##    Country      Age Salary Purchased
## 1   France 44.00000  72000        No
## 2    Spain 27.00000  48000       Yes
## 3  Germany 30.00000  54000        No
## 4    Spain 38.00000  61000        No
## 5  Germany 40.00000     NA       Yes
## 6   France 35.00000  58000       Yes
## 7    Spain 38.77778  52000        No
## 8   France 48.00000  79000       Yes
## 9  Germany 50.00000  83000        No
## 10  France 37.00000  67000       Yes
```

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

```r
#hot encoding
library(dummies)
```

```
## dummies-1.5.6 provided by Decision Patterns
```

```r
df <- dummy.data.frame(dataset, names=c("Country"), sep="_")
```

```
## Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts =
## FALSE): non-list contrasts argument ignored
```

```r
df
```

```
##    Country_France Country_Germany Country_Spain      Age   Salary
## 1               1               0             0 44.00000 72000.00
## 2               0               0             1 27.00000 48000.00
## 3               0               1             0 30.00000 54000.00
## 4               0               0             1 38.00000 61000.00
## 5               0               1             0 40.00000 63777.78
## 6               1               0             0 35.00000 58000.00
## 7               0               0             1 38.77778 52000.00
## 8               1               0             0 48.00000 79000.00
## 9               0               1             0 50.00000 83000.00
## 10              1               0             0 37.00000 67000.00
##    Purchased
## 1         No
## 2        Yes
## 3         No
## 4         No
## 5        Yes
## 6        Yes
## 7         No
## 8        Yes
## 9         No
## 10       Yes
```

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

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

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

```
##    Country      Age   Salary Purchased
## 1        1 44.00000 72000.00         0
## 2        2 27.00000 48000.00         1
## 3        3 30.00000 54000.00         0
## 4        2 38.00000 61000.00         0
## 5        3 40.00000 63777.78         1
## 6        1 35.00000 58000.00         1
## 7        2 38.77778 52000.00         0
## 8        1 48.00000 79000.00         1
## 9        3 50.00000 83000.00         0
## 10       1 37.00000 67000.00         1
```