wk_4_hw

Mrinal Vashisth

7 April 2019

```
######################
# Data Preprocessing #
######################

# Let's make a dataframe

Country <- c('France','Spain','Germany','Spain','Germany','France','Spain','France','Germany','France')
Age <- c('44','27','30','38','40','35','','48','50','37')
Salary <- c('72000','48000','54000','61000','','58000','52000','79000','83000','67000')

dataset <- data.frame(Country, Age, Salary)
dataset
```

```
##    Country Age Salary
## 1   France  44  72000
## 2    Spain  27  48000
## 3  Germany  30  54000
## 4    Spain  38  61000
## 5  Germany  40       
## 6   France  35  58000
## 7    Spain      52000
## 8   France  48  79000
## 9  Germany  50  83000
## 10  France  37  67000
```

```
"We can see that dataset is with missing values."
```

```
## [1] "We can see that dataset is with missing values."
```

```
# Taking care of missing data

# Although there are many options to take care of missing values, we are going to use 'replace by Mean approch'

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)

dataset
```

```
##    Country Age Salary
## 1   France   8      8
## 2    Spain   2      2
## 3  Germany   3      4
## 4    Spain   6      6
## 5  Germany   7      1
## 6   France   4      5
## 7    Spain   1      3
## 8   France   9      9
## 9  Germany  10     10
## 10  France   5      7
```

```
"Here dataset ain't got no missing values."
```

```
## [1] "Here dataset ain't got no missing values."
```