# Biomedical Data Science - Assignment 1

*Damian Baeza*

*14 February 2018*

## Problem 1

### Question (a)

```
# Data frame with the airquality data
df.a <- airquality

# Number of missing values in Ozone variable
Ozone.Miss <- sum(is.na(df.a$Ozone))
print(Ozone.Miss)
```

```
## [1] 37
```

```
# Mean imputation of Ozone variable
df.a$Ozone[is.na(df.a$Ozone)] <- mean(df.a$Ozone, na.rm = TRUE)

# Number of missing values in Ozone variable after mean imputation
Ozone.Miss <- sum(is.na(df.a$Ozone))
print(Ozone.Miss)
```

```
## [1] 0
```

### Question (b)

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
impute.to.window.mean <- function(x, windowsize){
  ## Function that imputes the mean of a vector considering a windowsize.
  ## A missing value will be imputed with the mean of the values in the
  ## positions (i - windowsize) and (i + windowsize).
  ## Inputs: vector to be imputed and size of window
  ## Outputs: imputed vector
  if(windowsize > length(x)){
    warning("Window size is longer than vector, vector length will be used instead")
    impute.val <- mean(x, na.rm = TRUE)
    x[is.na(x)] <- impute.val
    return(x)
  }
  else if(windowsize < 0){
    stop("Window size is negative!")
  }
  else{
    # position of missing values
    Miss.pos <- which(is.na(x))
```

```
    new.x <- x
    for(i in Miss.pos){
      # calculation of begining of the window of values used for
      # imputation
      first <- max(i - windowsize, 1)
      # calculation of end of the window of values used for
      # imputation
      last <- min(i + windowsize, length(x))
      # mean calculation
      impute.val <- mean(x[first:last], na.rm = TRUE)
      # imputation of missing value
      new.x[i] <- impute.val
    }
    return(new.x)
  }
}
```

## Question (c)

```
df.c <- airquality$Ozone

n <- length(df.c)

windows.size <- c(10, 25, 50, 75, 100, 125)

Results.Q1 <- data.frame(n10 = numeric(n),
                         n25 = numeric(n),
                         n50 = numeric(n),
                         n75 = numeric(n),
                         n100 = numeric(n),
                         n125 = numeric(n))

for (i in 1:length(windows.size)) {
  Results.Q1[,i] <- impute.to.window.mean(x = df.c,
                                          windowsize = windows.size[i])
}

Abs.Diff <- round(abs(Results.Q1 - df.a$Ozone), 3)

colnames(Abs.Diff) <- c()

Max.Abs.Diff <- data.frame(Max.Abs.Diff = sapply(Abs.Diff, max), Window.Size = windows.size)

print(Max.Abs.Diff)
```

```
##   Max.Abs.Diff Window.Size
## 1       26.771          10
## 2       19.553          25
## 3       17.015          50
## 4        6.404          75
## 5        6.461         100
## 6        5.669         125
```