RR-Project-1

# Reproducible Research --- Project 1

## Analysis with NA ingnored

* Setup figure diretory

knitr::opts\_chunk$set(fig.path="figures/")

* Load data

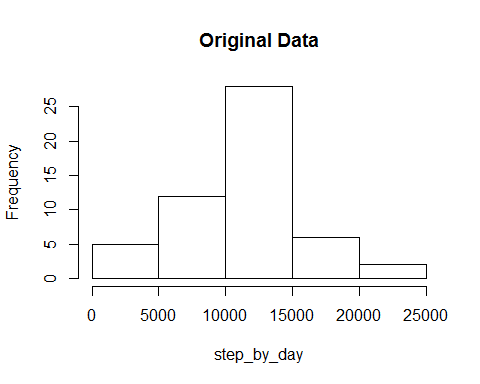
data <- read.csv(file="activity.csv", header = TRUE)

* Calculate total steps of each day

step\_by\_day <- with(data, tapply(steps, date, sum))

* Create a Histgram

orig\_hist <- hist(step\_by\_day, main = "Original Data")



## for comparison with new data set below  
 orig\_hist$count

## [1] 5 12 28 6 2

* Calculate mean and median, ignoring NAs

mean(step\_by\_day, na.rm = TRUE)

## [1] 10766.19

median(step\_by\_day, na.rm = TRUE)

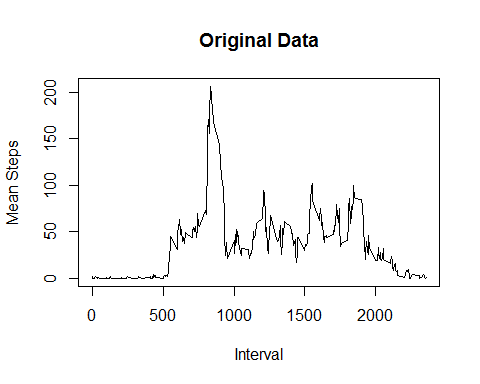
## [1] 10765

* Calculate mean by interval

mean\_by\_interval <- with(data, tapply(steps, interval, mean, na.rm = TRUE))

* Create a time series plot of mean steps averaged across all days

t\_series <- subset(data, date == "2012-10-01")$interval  
 plot(t\_series, mean\_by\_interval, type = "l", xlab="Interval", ylab = "Mean Steps", main = "Original Data")



* find the maximum

max(mean\_by\_interval)

## [1] 206.1698

which(mean\_by\_interval == max(mean\_by\_interval))

## 835   
## 104

## Imputing missing values: replacing NAs with the mean of the interval

* Calculate and report the total number of missing values in the dataset (i.e. the total number of rows with NAs)

sum(is.na(data$step))

## [1] 2304

**Discusion:** 2304 out of 17568 observations missing, about 13%

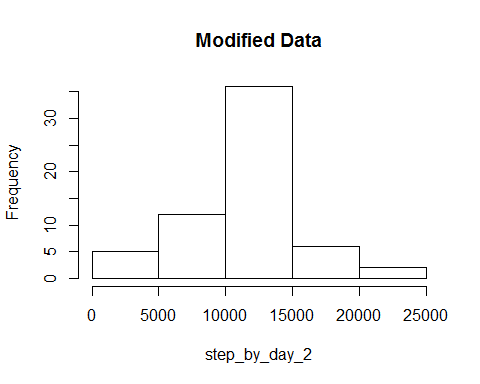
* use Mean of the interval to replace NA

num\_of\_dates <- nrow(step\_by\_day)  
 num\_0f\_interval <- nrow(mean\_by\_interval)  
   
   
 new\_data <- data.frame (steps = numeric(), date = factor(), interval = integer() )   
 d1 <- as.Date("2012-10-01")  
 for (i in 1:num\_of\_dates) {  
 d <- as.Date(d1 + i - 1)  
 tmp\_data <- subset(data, date == as.character(d))  
   
   
 for (j in 1:num\_0f\_interval) {  
 steps <- tmp\_data[j,1]  
 if (is.na(steps)) {  
 steps <- mean\_by\_interval[j]  
 }  
   
   
 if(!exists("new\_data")) {  
 new\_data <- c(steps, d, t\_series[j])  
 }  
 else {  
 tmp\_row <- c(steps, d, t\_series[j])  
 new\_data <- rbind(new\_data, tmp\_row)  
 }  
 }   
   
 }  
   
 colnames(new\_data) <- c("steps","date","interval")  
 new\_data$date <- as.Date(new\_data$date, origin ="1970-01-01")

## Statistics with new data set (NAs replaced by mean of the interval)

* Calculate total steps of each day and Plot a histgram

step\_by\_day\_2 <- with(new\_data, tapply(steps, date, sum))  
   
 new\_hist <- hist(step\_by\_day\_2, main = "Modified Data")



## for comparison with original data set  
 new\_hist$count

## [1] 5 12 36 6 2

* Calculate mean and median,

mean(step\_by\_day\_2, na.rm = TRUE)

## [1] 10766.19

median(step\_by\_day\_2, na.rm = TRUE)

## [1] 10766.19

**Discussion:** After imputing NAs, median and mean steps did not change with any significance, most likely due to the fact that NAs were replaced by mean of the interval. However, the peak of the histagram with steps between 10000 and 15000, increased significantly, while other bins have significant change. This deserves further investigation.

* Add a new column, wday, two level factor, weekday or weekend

weekday <- c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday")  
 wday\_vector <- factor((weekdays(as.Date(new\_data$date)) %in% weekday), levels=c(TRUE, FALSE), labels=c("weekday", "weekend"))  
 new\_data$wday <- wday\_vector

* Calculate mean by interal for weekday and weekend
* Make a panel plot of mean steps ~ interval for weekday and weekend

mean\_by\_interval\_wday <- with(new\_data, tapply(steps, list(wday, interval), mean))  
  
 par(mfrow = c(1,2), mar = c(5, 5, 4, 2))  
 plot(t\_series, mean\_by\_interval\_wday["weekday",], type ="l", ylim = c(0, 240), xlab = "Interval", ylab = "Mean Steps", main = "Weekday")  
 plot(t\_series, mean\_by\_interval\_wday["weekend",], type ="l", ylim = c(0, 240), xlab = "Interval", ylab = "Mean Steps", main = "Weekend")

