Activity

Read "activity.csv"

```
getwd()
```

```
## [1] "/Users/Hung/Desktop/data"
```

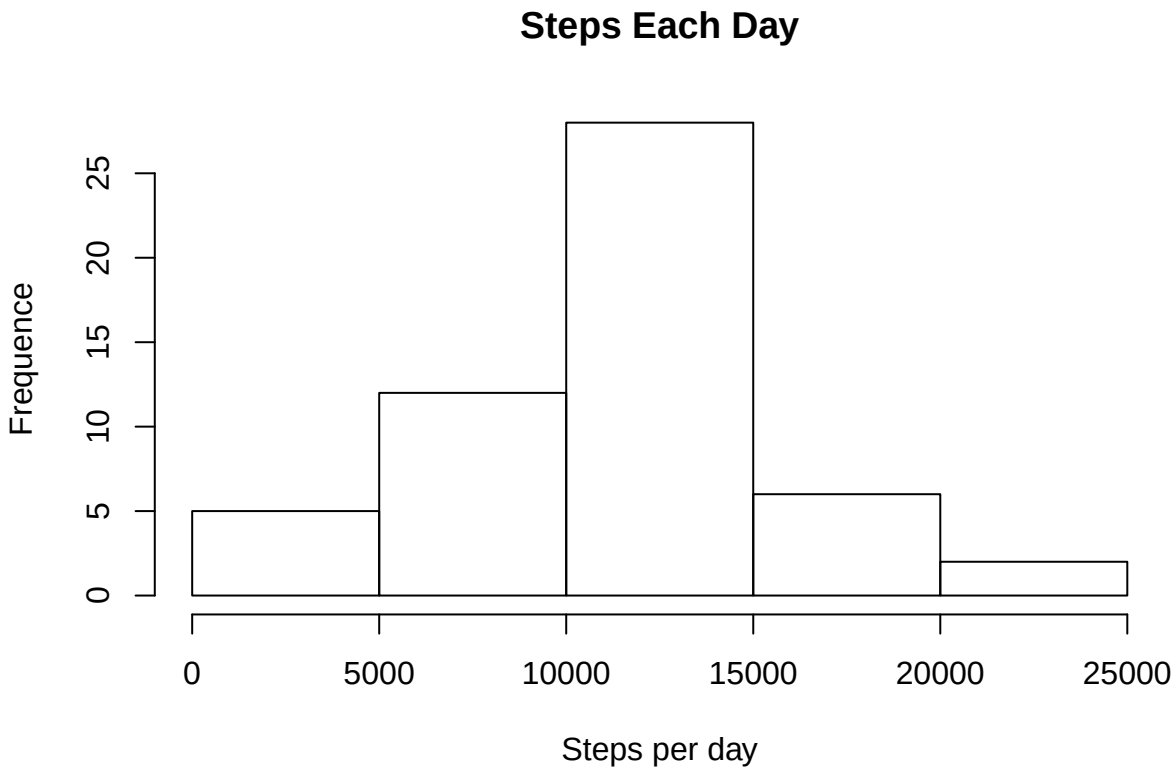
```
activity <- read.csv("./activity.csv")
```

Processing the data

```
activity$datetime <- as.POSIXct(activity$date,format = "%Y-%m-%d")
activity$day<- weekdays(as.Date(activity$date))
## clean the data
clean<- activity[!is.na(activity$steps),]
```

What is mean total number of steps taken per day?

```
sum <- tapply(clean$steps,clean$datetime, sum)
hist(sum, xlab = "Steps per day", ylab = "Frequence", main = "Steps Each Day")
```



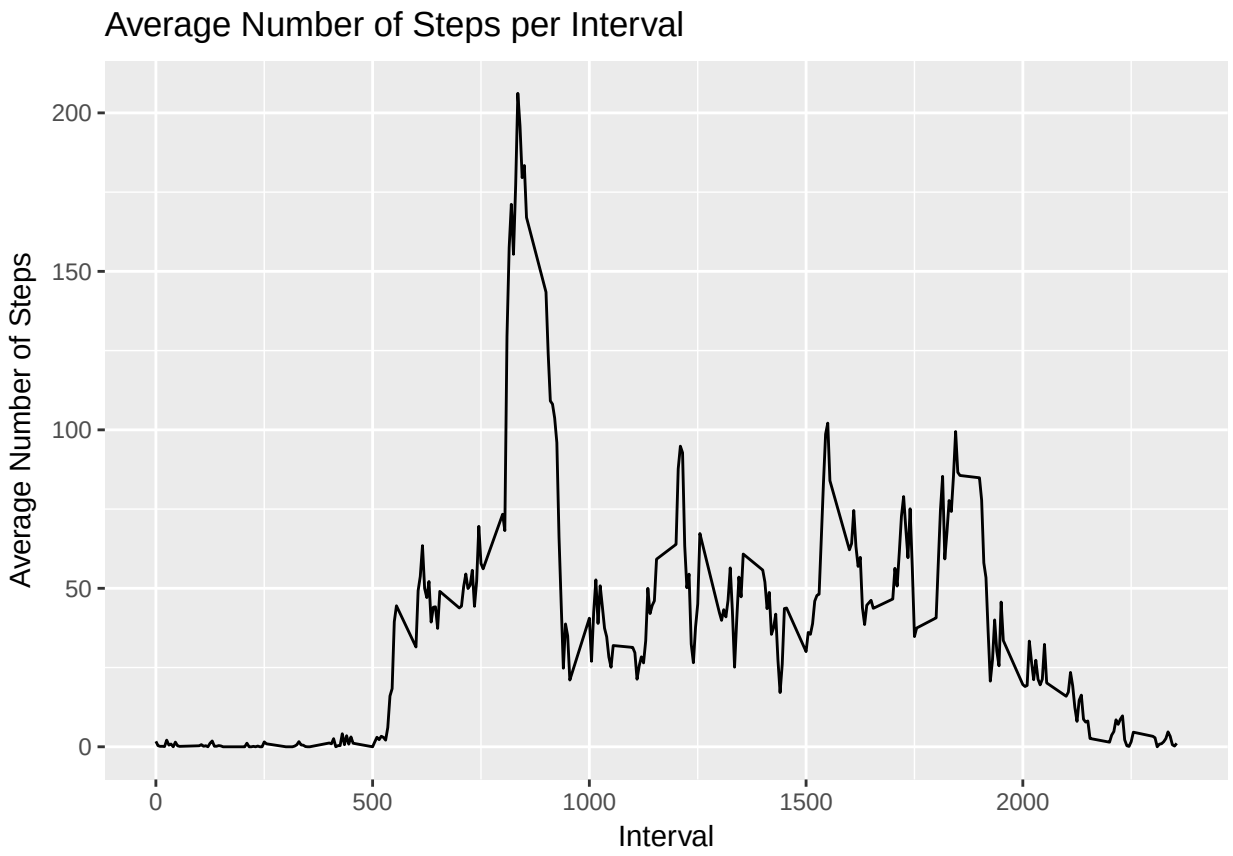

```
## Calculate and report the mean and median of the total number of steps taken per day
mean <- as.integer(mean(sum))
median <- as.integer(median(sum))
```

The average number of steps taken each day was 10766 steps. The median number of steps taken each day was 10765 steps.

What is the average daily activity pattern?

Make a time series plot (i.e. type = "l" of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all days (y-axis)

```
library(plyr)
library(ggplot2)
meaninterval <- ddply(clean,.(interval), .fun = summarize, Avg = mean(steps))
g <- ggplot(meaninterval, aes(x = interval, y = Avg), type = "l")
g+ geom_line() + xlab("Interval")+ylab("Average Number of Steps")+ggtitle("Average Number of Steps per I
```



Which 5-minute interval, on average across all the days in the dataset, contains the maximum number of steps?

```
## maximum number of steps
maxsteps <- max(meaninterval$Avg)
meaninterval[meaninterval$Avg == maxsteps,]
```

```
##     interval      Avg
## 104      835 206.1698
```

The maximum steps per interval is 206 steps. The 835 interval has the max number of steps.

Imputing missing values

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

```
##Number of NAs in original data set
nrow(activity[is.na(activity$steps),])
```

```
## [1] 2304
```

The total number of NA is 2304.

Devise a strategy for filling in all of the missing values in the dataset. The strategy does not need to be sophisticated. For example, you could use the mean/median for that day, or the mean for that 5-minute interval, etc.

My strategy is using the mean of that weekdays to fill in the missing values

```
## create the average number of steps per weekday and interval
new <- ddply(clean, .(interval, day), summarize, avg = mean(steps))
## create datasets with all NAs
ndata <- activity[is.na(activity$steps),]
## merge the dataset
merge <- merge(ndata, new, by = c("interval","day"))
```

Create a new dataset that is equal to the original dataset with missing values

```
## reorganzie the new dataset as the same format as the original dataset
new2 <- merge[,c(6,4,1,5,2)]
colnames(new2) <- c("steps","date","interval","datetime","day")
## merge the NA and non NA
merge2 <- rbind(clean, new2)
```

Make a histogram of the total number of steps taken each day and Calculate and report the mean and median total number of steps taken per day. Do these values differ from the estimates from the first part of the assignment? What is the impact of imputing missing data on the estimates of the total daily number of steps?

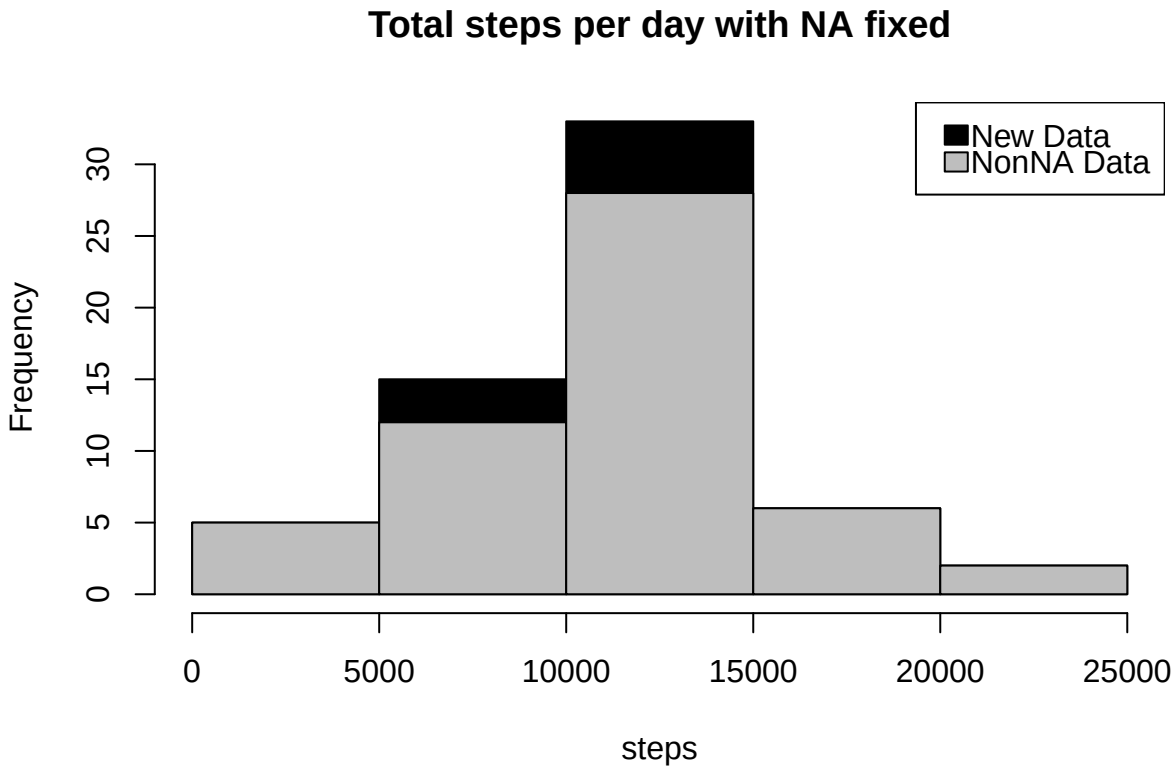
```
sum2 <-tapply(merge2$steps,merge2$datetime, sum)
as.integer(mean(sum2))
```

```
## [1] 10821
```

```
as.integer(median(sum2))
```

```
## [1] 11015
```

```
## creating the histogram with filled-in missing values
hist(sum2, xlab = "steps", main = "Total steps per day with NA fixed", col = "black")
hist(sum, col = "grey", add = T)
legend("topright", c("New Data", "NonNA Data"), fill=c("black", "grey"), y.intersp=0.7, x.intersp = 0.1)
```



The new mean total number of steps taken per day is 10821, compared to the data with Non NA 10766, there is a difference of 55 steps. The new median total number of steps taken per day is 11015, compared to the data with Non NA 10765, there is a difference of 250 steps. The overall shape of histogram doesn't change.

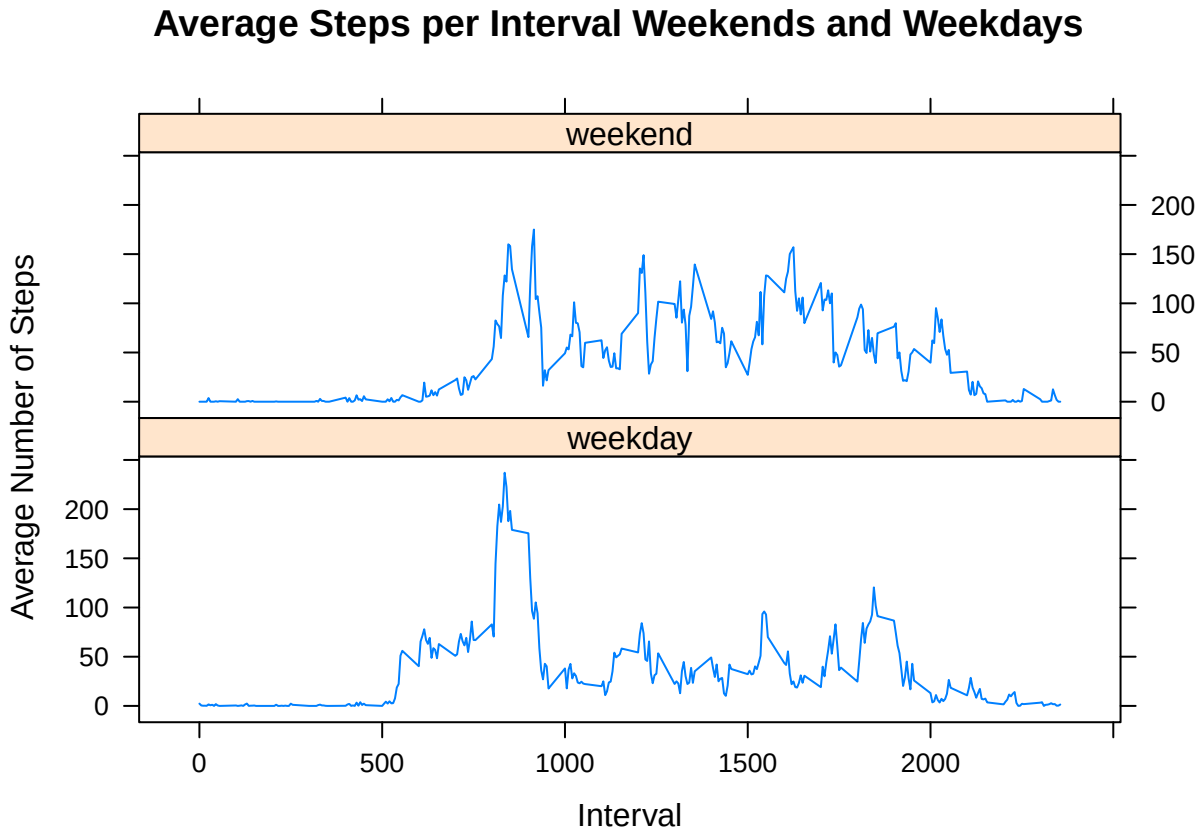
Are there differences in activity patterns between weekdays and weekends?

Create a new factor variable in the dataset with two levels – "weekday" and "weekend" indicating whether a given date is a weekday or weekend day.

```
merge2$daycategory <- ifelse(merge2$day %in% c("Saturday","Sunday"), "weekend", "weekday")
```

Make a panel plot containing a time series plot (i.e. type = "l") of the 5-minute interval (x-axis) and the average number of steps taken, averaged across all weekday days or weekend days (y-axis).

```
table <- ddply(merge2, .(interval, daycategory), summarize, avg = mean(steps))
library(lattice)
xyplot(avg ~ interval|daycategory, data = table, type = "l", layout = c(1,2),main="Average Steps per In
       ylab="Average Number of Steps", xlab="Interval")
```



There are differences of activities between weekends and weekdays.The activities during weekends are more spread out generally.