The Quantified Self

Getting the data

```
# Load the data
setwd("C:/repos_github/coursera/repres/data")
dt1 <- read.table("activity.csv", sep = ",", header = TRUE)
dt1 <- data.table(dt1)
#names(dt1)
#length(names(dt1))

# Remove missing values
dt2 <- dt1[complete.cases(dt1),]

# Rows with missing data removed
naRemoved <- nrow(dt1) - nrow(dt2)
dt1 <- dt2
remove(dt2)
```

Analysis part 1 - Histogram

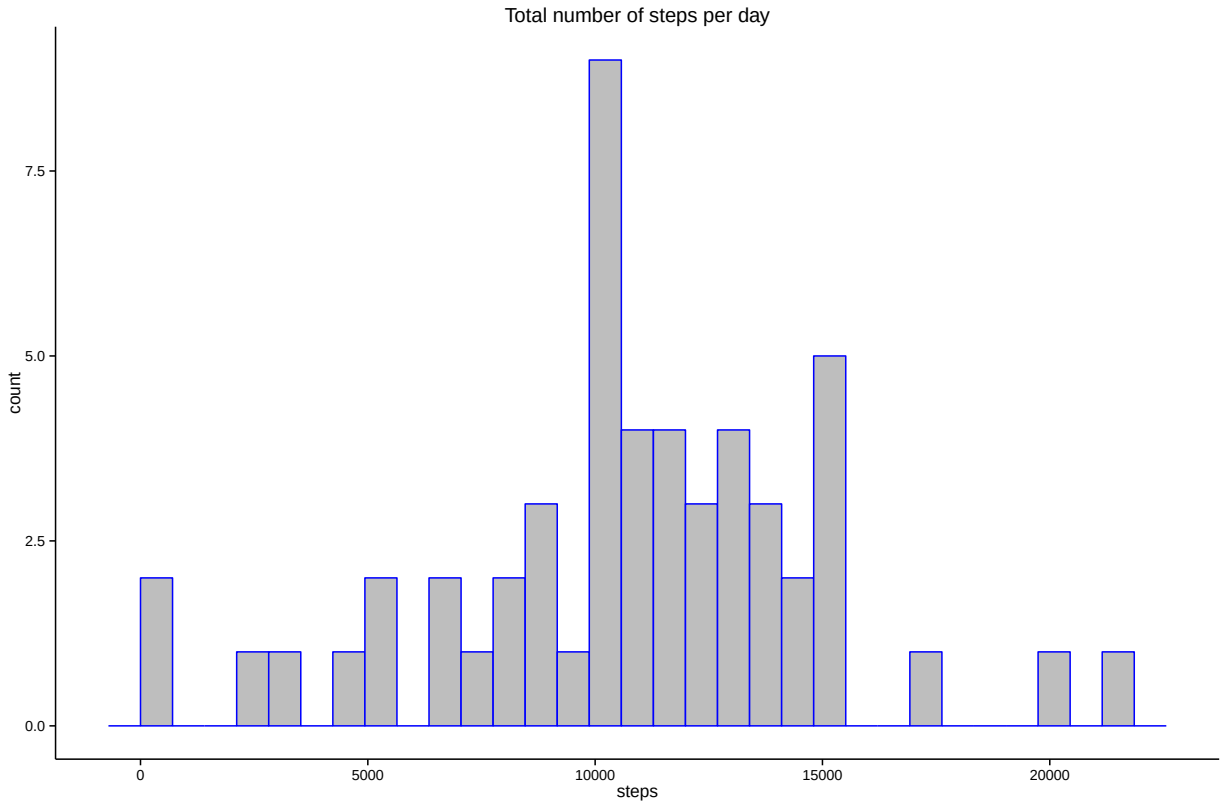
```
# Number of steps per day, data.table
dt2 <- dt1[,.(steps.sum = sum(steps)),by=date]

# Number of steps per day, sqldf
dt2s <- sqldf("SELECT sum(steps), date
        FROM dt1
        Group by date")

#names(dt2)

# Number of steps per day, calculations
nsteps <- sum(dt2$steps)
avgsteps <- mean(dt2$steps.sum)
medsteps <- median(dt2$steps.sum)


# Histogram of number of steps per day
h1 <- ggplot(data=dt2, aes(dt2$steps.sum)) + geom_histogram(colour = "blue", fill = "grey")
h1 <- h1 + theme_classic()
h1 <- h1 + ggtitle("Total number of steps per day") + xlab("steps")
plot(h1)
```




```
setwd("C:/repos_github/coursera/repres")
#ggsave(filename = "Histogram Number of Steps.pdf", plot = h1)
```

The average number of steps taken per day, without removing missing values, is 570608. The mean and median for the same category is 10788.19 and 10765.

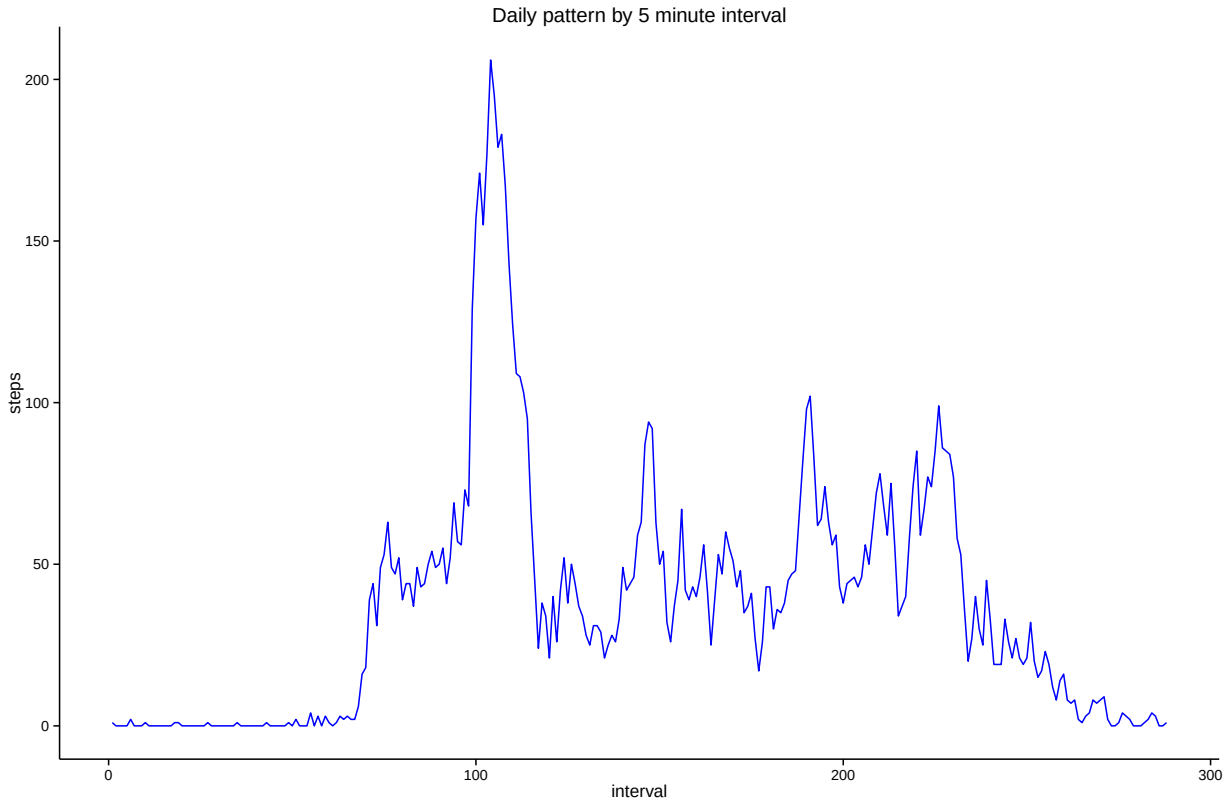
Analysis part 2 - Daily pattern

```
#   1   Average number of steps taken, averaged across all days
#   1.1 Data
dt3 <- sqldf("SELECT interval, avg(steps) as steps
        FROM dt1
        Group by interval")

idx1 <- 1:nrow(dt3)
dt4 <- data.table(dt3, idx1)

#   1.2 Plot
p1 <- ggplot(dt4, aes((interval), steps)) + geom_line(colour = "blue") + theme_classic() +
      ggtitle("Daily pattern by 5 minute interval") + xlab("interval")
#p1

p2 <- ggplot(dt4, aes((idx1), steps)) + geom_line(colour = "blue") + theme_classic() +
      ggtitle("Daily pattern by 5 minute interval") + xlab("interval")
p2
```




```
#   2   Which 5 minute interval has the max number of steps?
maxsteps = max(dt4$steps)
maxint <- sqldf("SELECT interval, max(steps)
        FROM dt4")
maxint$interval[1]
```

```
## [1] 835
```

```
#ggsave(filename = "Daily pattern.pdf", plot = p1)
```

The interval which contains the maximum average steps per day, is 835.