Reproduce_Project_2

Sudabe

April 16, 2016

First I read the data in r and name the file activity.

```
activity<-read.csv("./activity.csv", header = TRUE)
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(lubridate)
```

I named the file activity. I also need to use "dplyr" and "lubridate" libraries.

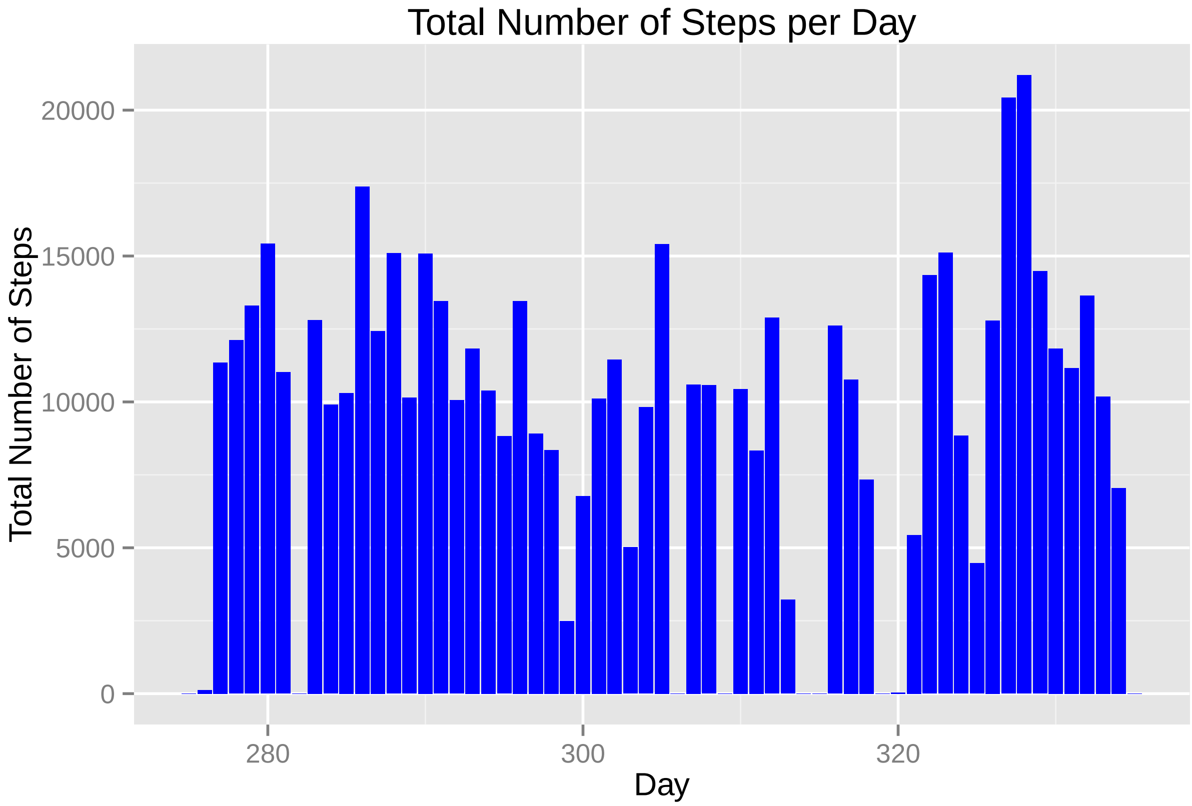
What is mean total number of steps taken per day?

Total number of steps:

```
activity$date=as.character(activity$date)
activity$date=as.Date(activity$date)
activity$day=yday(activity$date)
activity_day=group_by(activity, day)
Total_step_day <- summarise(activity_day, steps_day=sum(steps, na.rm = TRUE ))
```

Histogram:

```
library(ggplot2)
g=ggplot(Total_step_day, aes(day, steps_day))+geom_histogram(stat = "identity", fill="blue")+ggtitle("T
plot(g)
```



Calculate and report the mean and median of the total number of steps taken per day:

```
step_day_stat <- summarise(activity_day, mean_steps_day=mean(steps, na.rm = TRUE ), median_steps_day=me
step_day_stat
```

```
## Source: local data frame [61 x 3]
##
##      day mean_steps_day median_steps_day
##    (dbl)          (dbl)            (dbl)
## 1    275            NaN               NA
## 2    276        0.43750                0
## 3    277       39.41667                0
## 4    278       42.06944                0
## 5    279       46.15972                0
## 6    280       53.54167                0
## 7    281       38.24653                0
## 8    282            NaN               NA
## 9    283       44.48264                0
## 10   284       34.37500                0
## ..   ...            ...              ...
```

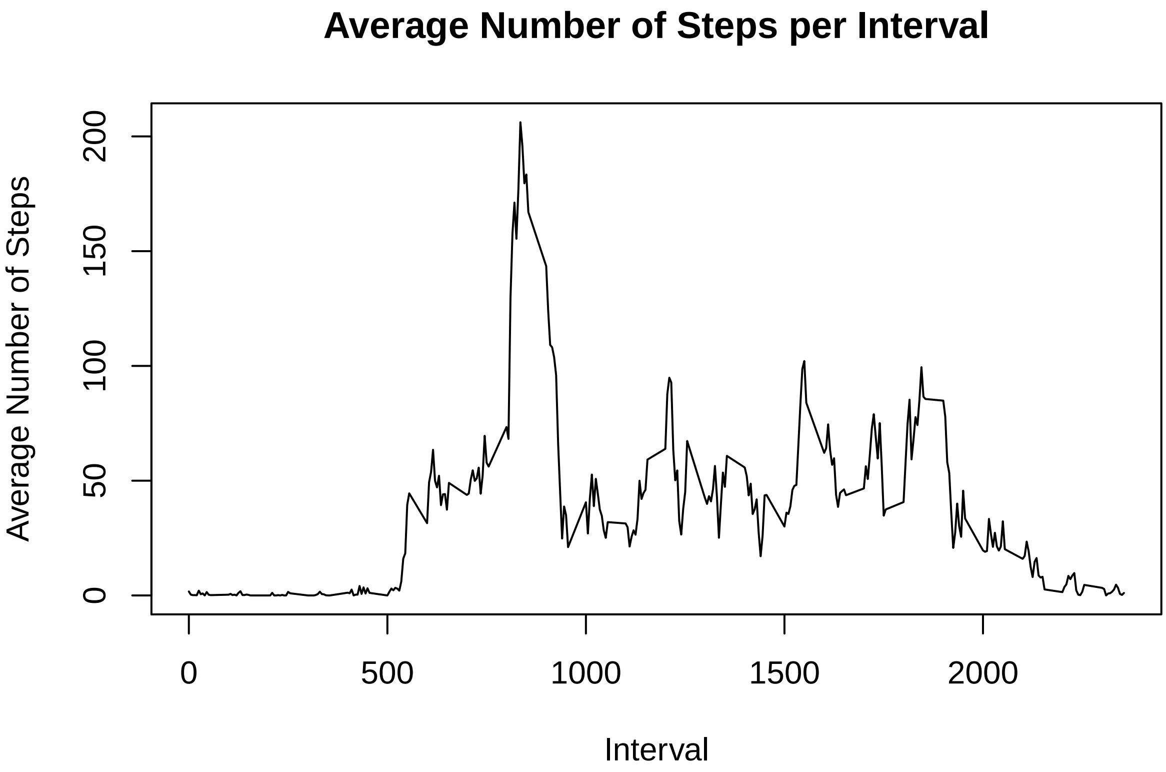
```
summarize(activity, mean=mean(steps, na.rm = TRUE), median=median(steps, na.rm = TRUE))
```

```
##      mean median
## 1 37.3826      0
```

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)

```
step_interval = group_by(activity, interval)
step_interval_mean= summarise(step_interval, mean=mean(steps, na.rm= T))
with(step_interval_mean, plot(interval, mean, type="l", main="Average Number of Steps per Interval", xla
```



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

```
which.max(step_interval_mean$mean)
```

```
## [1] 104
```

```
step_interval_mean[104,]
```

```
## Source: local data frame [1 x 2]
##
##   interval     mean
##      (int)    (dbl)
## 1      835 206.1698
```

So interval 835 has the maximum average 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)

```
sum(is.na(activity$steps))
```

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

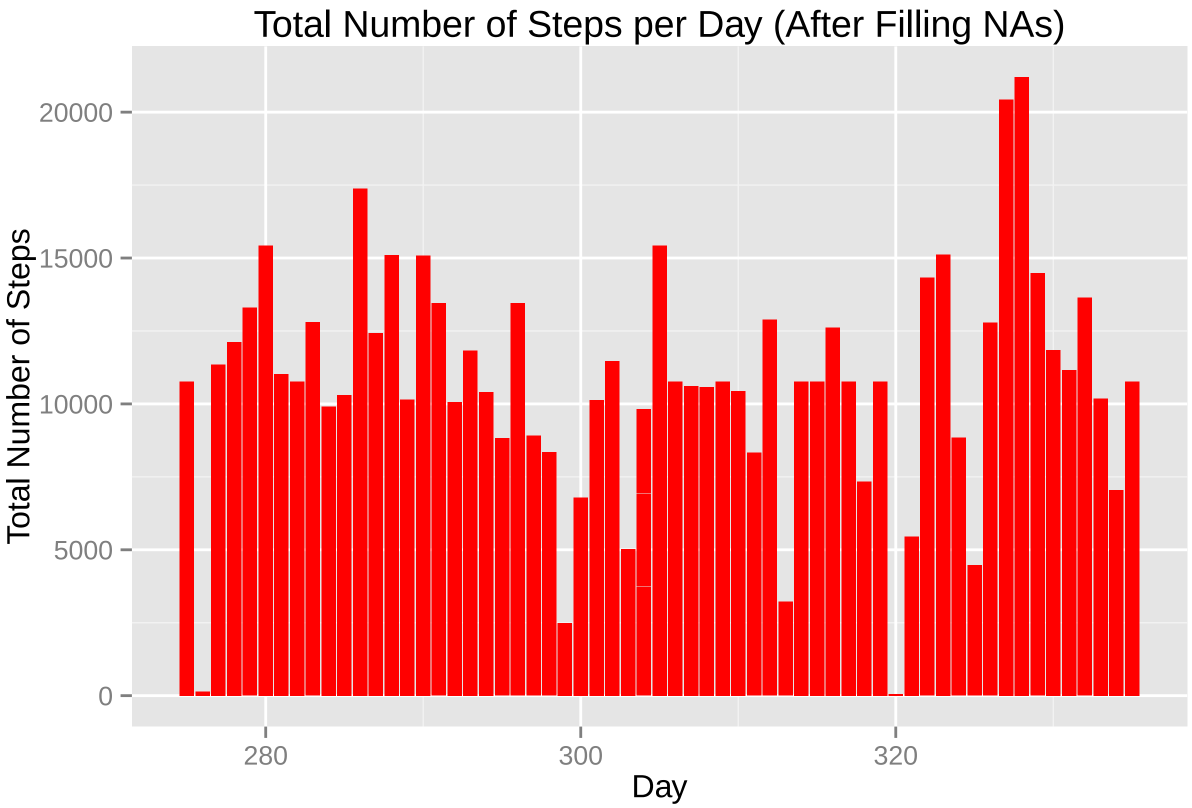
A strategy for filling in all of the missing values.

My strategy is to use the average of that day for the missing value, if the average of the day is missing then I use the overall average.

```
total_mean=mean(activity$steps, na.rm =TRUE)
activity_mean=merge(activity, step_day_stat, by=intersect(names(activity), names(step_day_stat)))
activity_mean2=merge(activity, step_day_stat, by=intersect(names(activity), names(step_day_stat)))

for(i in 1:length(activity_mean$steps)){if(is.na(activity_mean[i,5])){activity_mean[i,5]=total_mean}}
for(i in 1:length(activity_mean$steps)){if(is.na(activity_mean[i,2])){activity_mean[i,2]=activity_mean[
```

```
g3=ggplot(activity_mean, aes(day, mean_steps_day))+geom_histogram(stat = "identity", fill="red")+ggtitl
plot(g3)
```



```
activity_mean_day=group_by(activity_mean, day)
summarize(activity_mean_day, mean=mean(steps), median=median(steps))
```

```
## Source: local data frame [61 x 3]
##
##      day     mean  median
##    (dbl)    (dbl)   (dbl)
## 1    275 37.38260 37.3826
## 2    276  0.43750  0.0000
## 3    277 39.41667  0.0000
## 4    278 42.06944  0.0000
## 5    279 46.15972  0.0000
## 6    280 53.54167  0.0000
## 7    281 38.24653  0.0000
## 8    282 37.38260 37.3826
## 9    283 44.48264  0.0000
## 10   284 34.37500  0.0000
## ..   ...      ...     ...
```

```
summarize(activity_mean, mean=mean(steps), median=median(steps))
```

```
##      mean median
## 1 37.3826      0
```

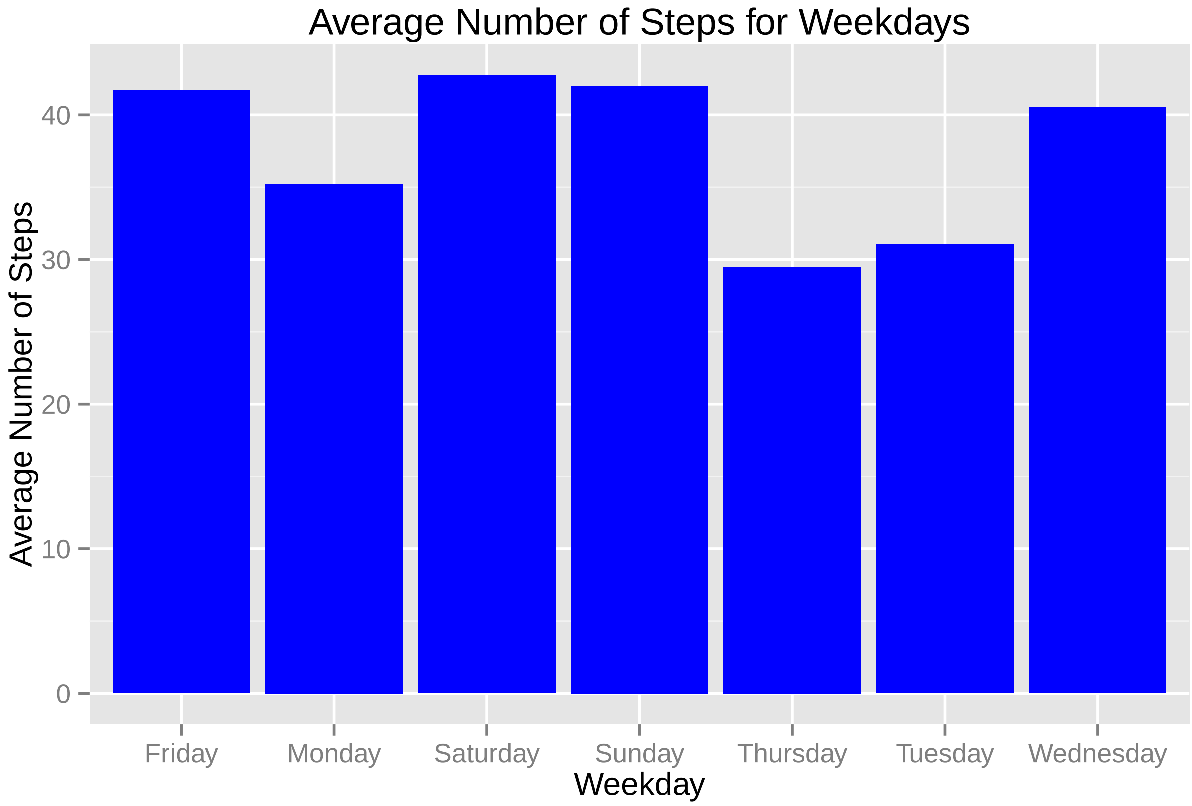
Total mean is not changing because my analysis is showing that either the value for all intervals in a day is missing or non of them is missing. This means that with the strategy that I used all missing values are substituted by total mean of the data which is 37.3826. Because of this the total mean after filling the missing values won't change.If you compare figure . . . you will see that the messing values are substituted by the average total step in a day which is :37.3826*288= 10,656

Are there differences in activity patterns between weekdays and weekends?

Create a new factor variable in the dataset with two levels - "weekday" and "weekend"

For this part I used the filled data from previous part

```
activity_mean$wday=weekdays(activity_mean$date)
activity_wday=group_by(activity_mean, wday)
activity_wday_mean=summarize(activity_wday, mean_wday=mean(steps, na.rm =TRUE))
g4=ggplot(activity_wday_mean, aes(wday, mean_wday))+geom_histogram(stat = "identity", fill="blue")+ggti
plot(g4)
```

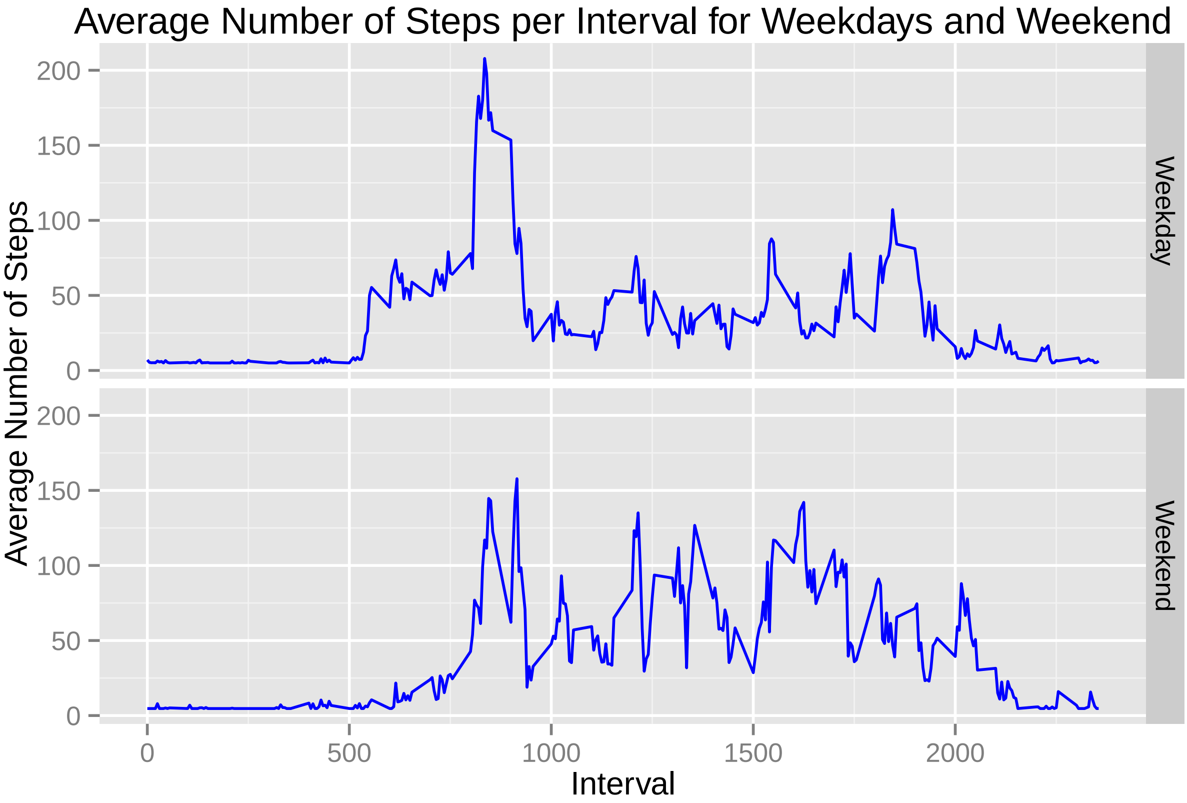


```
activity_mean$ffday="Weekday"
for(j in 1:length(activity_mean$steps)){if(activity_mean$wday[j]=="Saturday"|activity_mean$wday[j]=="Su
```

We can see that the average number of steps is generally higher during Weekend.

Make a panel plot

```
activity_interval=group_by(activity_mean, ffday, interval)
activity_interval_mean=summarize(activity_interval, mean_interval=mean(steps, na.rm =TRUE))
ggplot(activity_interval_mean, aes(interval, mean_interval))+geom_line(col="blue")+facet_grid(ffday~.)+g
```



By comparing number of steps in each interval we can see that it is not the same during weekend and weekdays. However there is no specific trend sometimes it is higher in weekends and sometimes it is lower.