Reproducible Research_P1

Santi

6/7/2562

Reproducible Research_P1

Peer-graded Assignment: Course Project 1 Skip to content

Search or jump to…

Pull requests Issues Marketplace Explore

@santimai 18 76 31,876 rdpeng/RepData_PeerAssessment1 Code Issues 5 Pull requests 266 Projects 0 Wiki Security Insights Peer Assessment 1 for Reproducible Research 13 commits 1 branch 0 releases 3 contributors @rdpeng rdpeng figures –> figure 1 Latest commit 80edf39 on Oct 22, 2014 Type Name Latest commit message Commit time doc Update instructions to add SHA-1 hash 6 years ago instructions_fig Update instructions/README 6 years ago PA1_template.Rmd Added “keep_md” YAML to template 5 years ago README.md figures –> figure 5 years ago activity.zip Add activity dataset 6 years ago README.md Introduction It is now possible to collect a large amount of data about personal movement using activity monitoring devices such as a Fitbit, Nike Fuelband, or Jawbone Up. These type of devices are part of the “quantified self” movement – a group of enthusiasts who take measurements about themselves regularly to improve their health, to find patterns in their behavior, or because they are tech geeks. But these data remain under-utilized both because the raw data are hard to obtain and there is a lack of statistical methods and software for processing and interpreting the data.

This assignment makes use of data from a personal activity monitoring device. This device collects data at 5 minute intervals through out the day. The data consists of two months of data from an anonymous individual collected during the months of October and November, 2012 and include the number of steps taken in 5 minute intervals each day.

```
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(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
## The following object is masked from 'package:base':
## 
##     date
```

```
library(stringr)
library(RODBC)
library(ggfortify)#autoplot
#1.Code for reading in the dataset and/or processing the data
setwd("C:\\Users\\CPUser\\Desktop\\Data analytics\\C5_Reproducible Research\\repdata_data_act
ivity")
#read data fram
df <- read.csv("activity.csv")
str(df)
```

```
## 'data.frame':    17568 obs. of  3 variables:
##  $ steps   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ date    : Factor w/ 61 levels "2012-10-01","2012-10-02",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ interval: int  0 5 10 15 20 25 30 35 40 45 ...
```

##Convent Factor to date time

```
df$date <- as.POSIXct(df$date, format="%Y-%m-%d")
range(df$date)
```

```
## [1] "2012-10-01 +07" "2012-11-30 +07"
```

```
str(df)
```

```
## 'data.frame':    17568 obs. of  3 variables:
##  $ steps   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ date    : POSIXct, format: "2012-10-01" "2012-10-01" ...
##  $ interval: int  0 5 10 15 20 25 30 35 40 45 ...
```

```
summary(df)
```

```
##      steps             date                interval     
##  Min.   :  0.00   Min.   :2012-10-01   Min.   :   0.0  
##  1st Qu.:  0.00   1st Qu.:2012-10-16   1st Qu.: 588.8  
##  Median :  0.00   Median :2012-10-31   Median :1177.5  
##  Mean   : 37.38   Mean   :2012-10-31   Mean   :1177.5  
##  3rd Qu.: 12.00   3rd Qu.:2012-11-15   3rd Qu.:1766.2  
##  Max.   :806.00   Max.   :2012-11-30   Max.   :2355.0  
##  NA's   :2304
```

```
dim(df)
```

```
## [1] 17568     3
```

Filler N/A

```
s_steps<-aggregate(df$steps,by=list(df$date),FUN=sum,na.rm=TRUE)
colnames(s_steps) <- c( "Date", "Step")
summary(s_steps)
```

```
##       Date                 Step
##  Min.   :2012-10-01   Min.   :    0
##  1st Qu.:2012-10-16   1st Qu.: 6778
##  Median :2012-10-31   Median :10395
##  Mean   :2012-10-31   Mean   : 9354
##  3rd Qu.:2012-11-15   3rd Qu.:12811
##  Max.   :2012-11-30   Max.   :21194
```

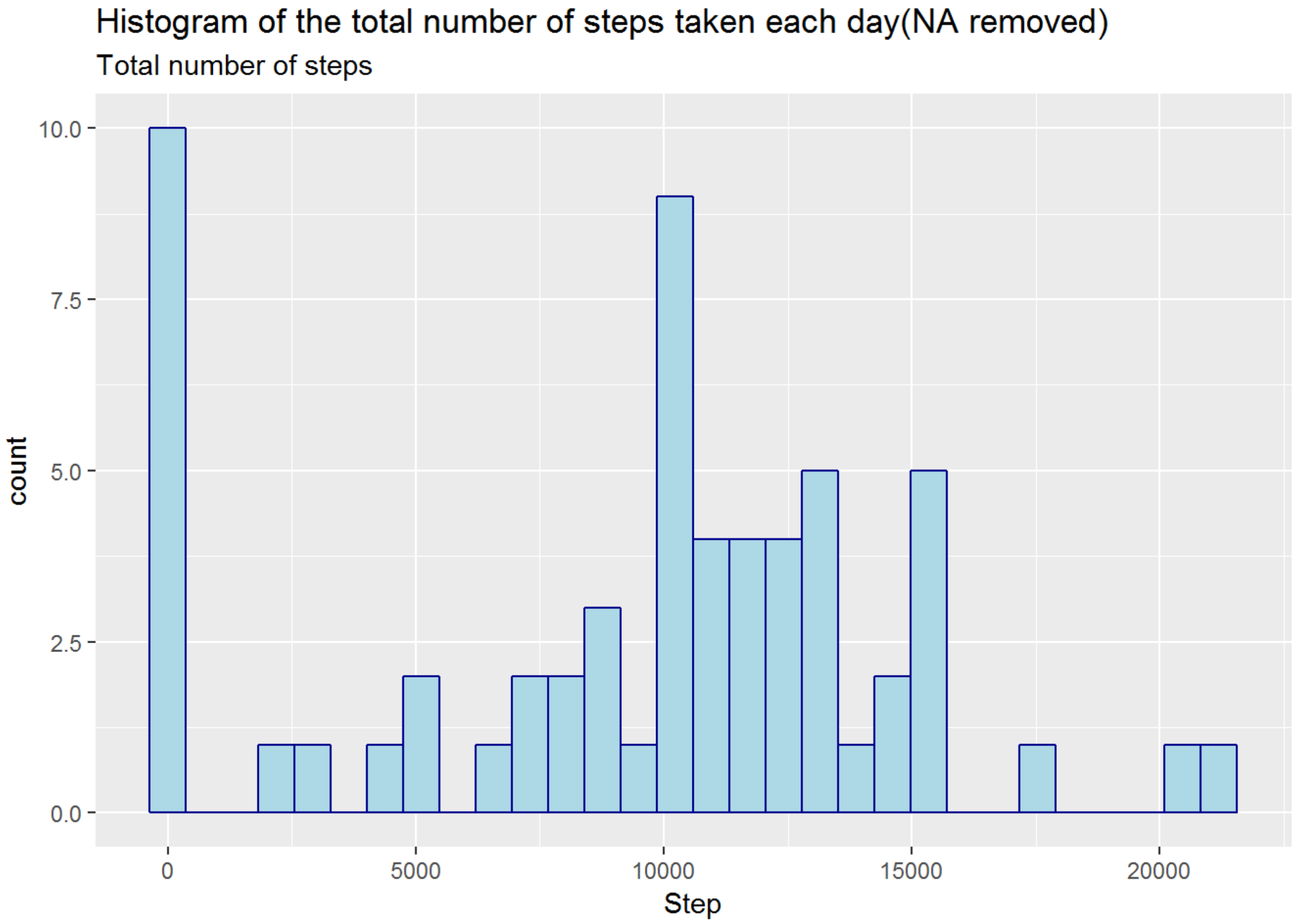
```
dim(s_steps)
```

```
## [1] 61  2
```

##2.Histogram of the total number of steps taken each day

```
S1 <- ggplot(s_steps, aes(x = Step )) + geom_histogram (color="darkblue", fill="lightblue")+
 labs(subtitle="Total number of steps",title = "Histogram of the total number of steps taken
 each day(NA removed)")
print(S1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



##3.Mean and median number of steps taken each day

```
mean(s_steps$Step)
```

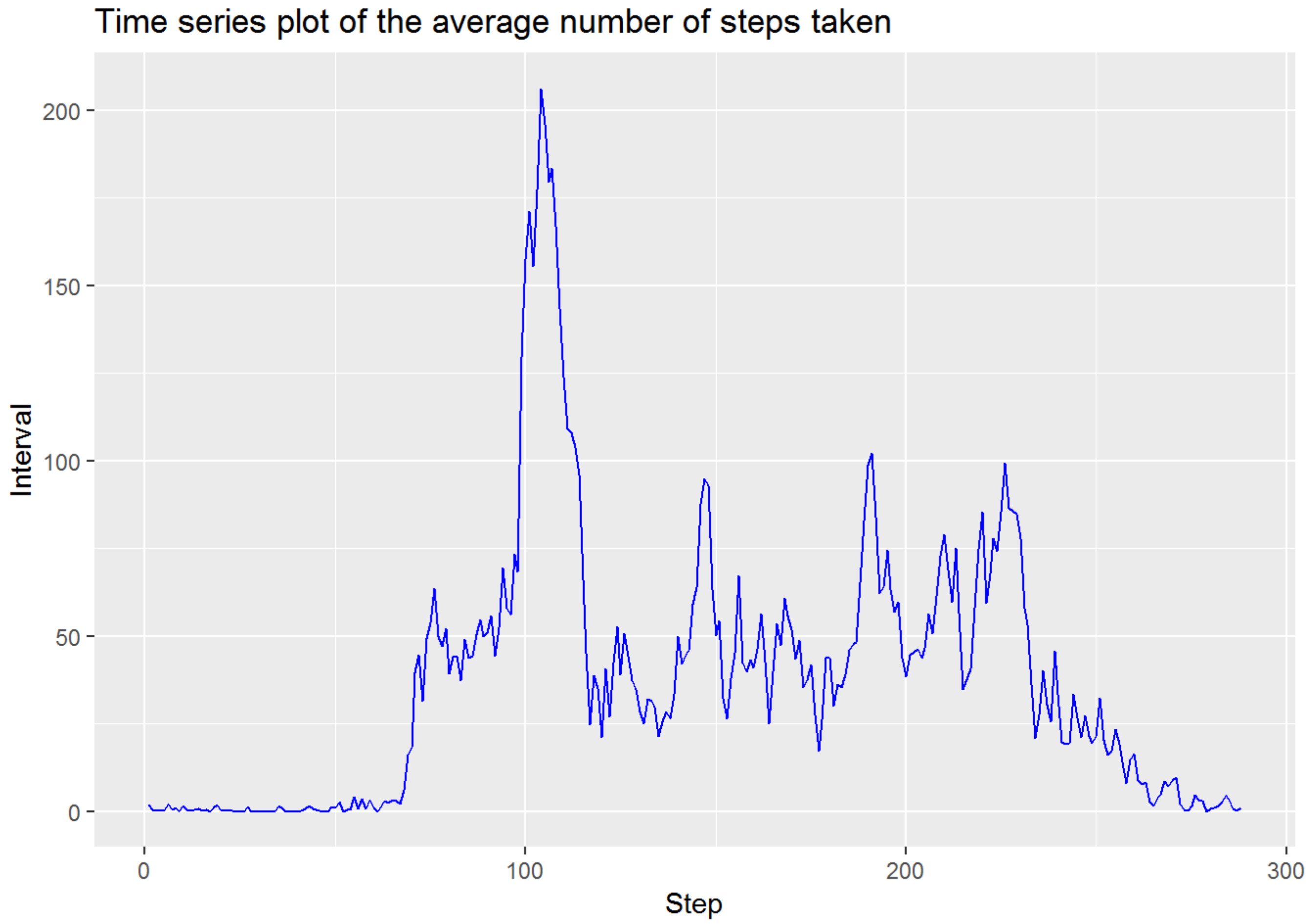
```
## [1] 9354.23
```

```
median(s_steps$Step)
```

```
## [1] 10395
```

##4.Time series plot of the average number of steps taken

```
avg_steps<-aggregate(df$steps,by=list(df$interval),FUN=mean,na.rm=TRUE)
colnames(avg_steps) <- c("interval","Step")
dt <- ts(avg_steps, start=1,frequency=1)
autoplot(dt[,"Step" ],colour = "blue")+
    ggtitle("Time series plot of the average number of steps taken ") +
    xlab("Step") +
    ylab("Interval")
```



##5.The 5-minute interval that, on average, contains the maximum number of steps

```
avg_steps[avg_steps$Step == max(avg_steps$Step),1]
```

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

##6.Code to describe and show a strategy for imputing missing data

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

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

```
df$steps[is.na(df$steps)] <- mean(df$steps,na.rm=TRUE)
sum(is.na(df$steps))
```

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

##7.Histogram with Repaced NA values

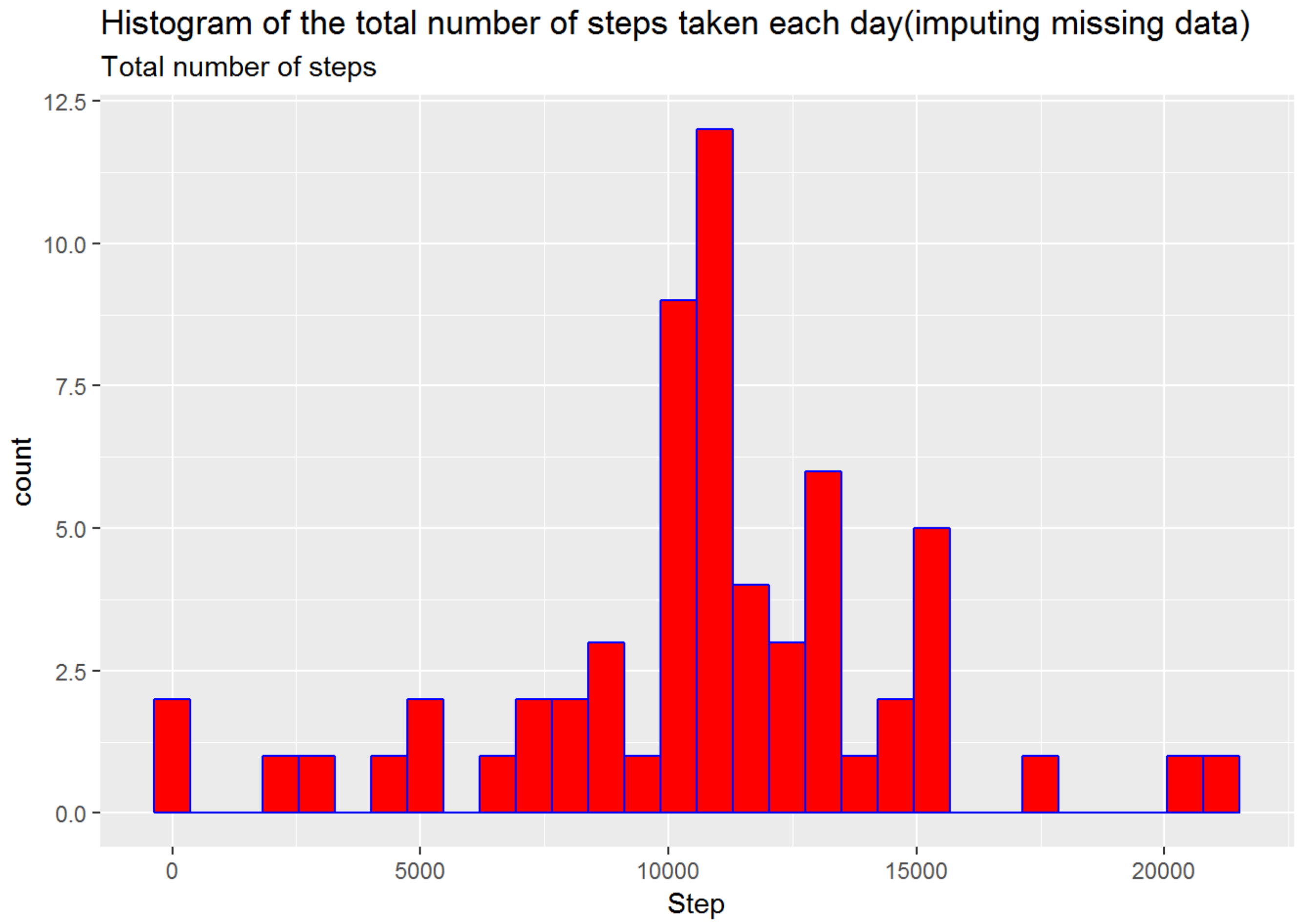
```
head(df)
```

	steps	date	interval
	<dbl>	<S3: POSIXct>	<int>
1	37.3826	2012-10-01	0
2	37.3826	2012-10-01	5
3	37.3826	2012-10-01	10
4	37.3826	2012-10-01	15
5	37.3826	2012-10-01	20
6	37.3826	2012-10-01	25

6 rows

```
s2_steps<-aggregate(df$steps,by=list(df$date),FUN=sum,na.rm=TRUE)
colnames(s2_steps) <- c( "Date", "Step")
S2 <- ggplot(s2_steps, aes(x = Step )) + geom_histogram (color = "blue", fill="red")+ labs(su
btitle="Total number of steps",title = "Histogram of the total number of steps taken each day
(imputing missing data)")
print(S2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
mean(s2_steps$Step)
```

```
## [1] 10766.19
```

```
median(s2_steps$Step)
```

```
## [1] 10766.19
```

##8.Panel plot comparing the average number of steps taken per 5-minute interval across weekdays and weekends

```
# Convert date into weekdays
df$days <- tolower(weekdays(df$date))
head(df)
```

	steps	date	interval	days
	<dbl>	<S3: POSIXct>	<int>	<chr>
1	37.3826	2012-10-01	0	จันทร์
2	37.3826	2012-10-01	5	จันทร์
3	37.3826	2012-10-01	10	จันทร์
4	37.3826	2012-10-01	15	จันทร์
5	37.3826	2012-10-01	20	จันทร์

	steps <dbl>	date <S3: POSIXct>	interval <int>	days <chr>
6	37.3826	2012-10-01	25	จันทร์

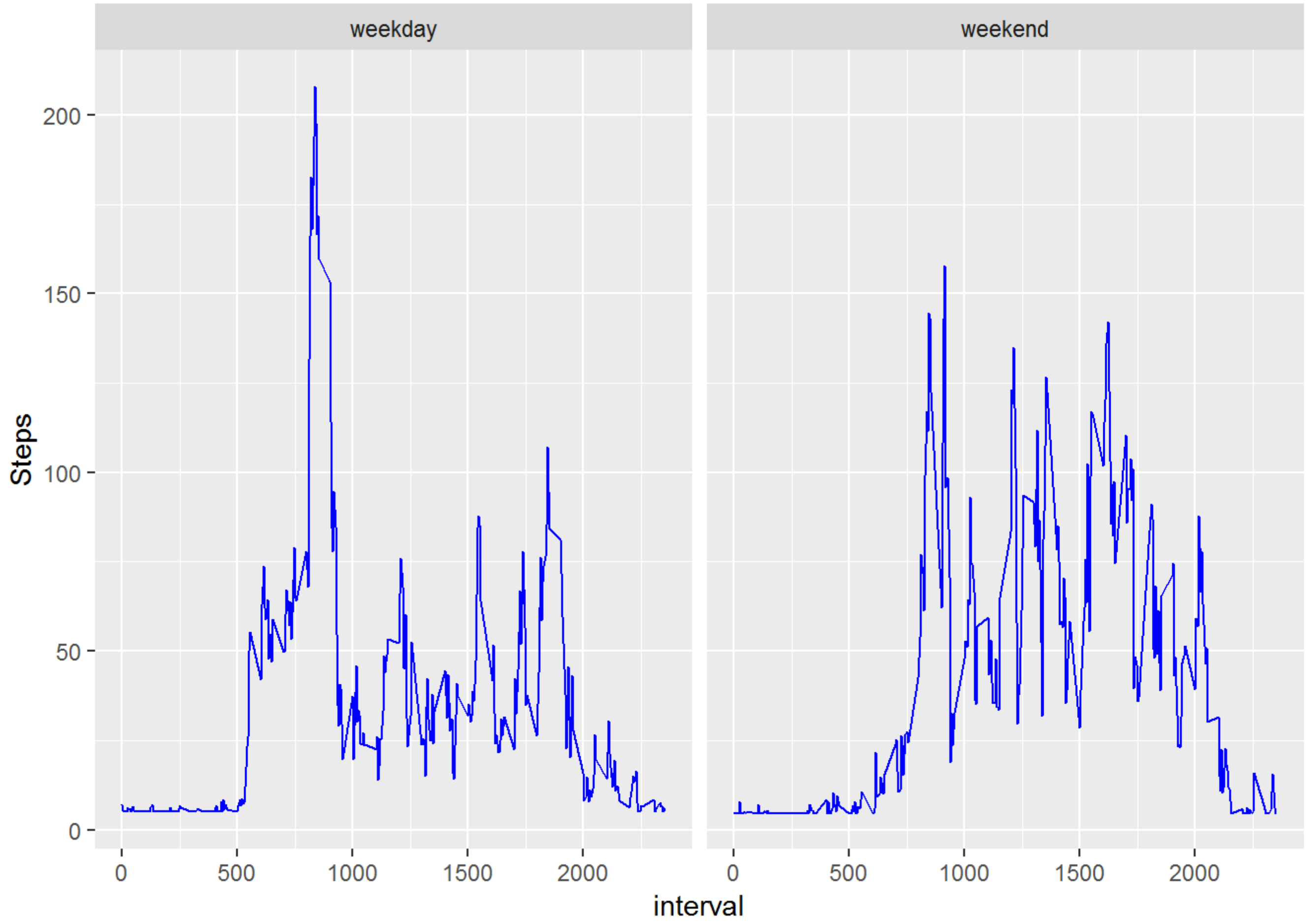
6 rows

```
df$day_type<-ifelse (df$days=="เสาร์" | df$day =="อาทิตย์","weekend","weekday")

#average number of steps taken per 5-minute
avg_steps<-aggregate(df$steps,by=list(df$interval,df$day_type),FUN=mean,na.rm=TRUE)
colnames(avg_steps) <- c("interval","day_type","Steps")

#plot comparing
S3 <- ggplot(aes(x=interval,y=Steps),data=avg_steps)+geom_line(color ="blue")+facet_wrap(~avg
_steps$day_type)

print(S3)
```


weekday
weekend
Steps
interval