Activity Analysis

Reproducable Research Week 2

Robin Smith

March 9, 2017

Loading and preprocessing the data

The activity monitoring data is downloaded from course website (https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip). The file is activity.zip and when unzipped activity.csv becomes available. This file is loaded into data variable d. A data table (dt) is created for processing with dplyr package.

```
download.file('https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip',
'activity.zip')
unzip ("activity.zip", exdir = "./")
d <- read.table('activity.csv', header = TRUE, sep=',')
dt <- tbl_df(d)
```

The file contains the following headings.

```
names(d)
```

```
## [1] "steps"    "date"     "interval"
```

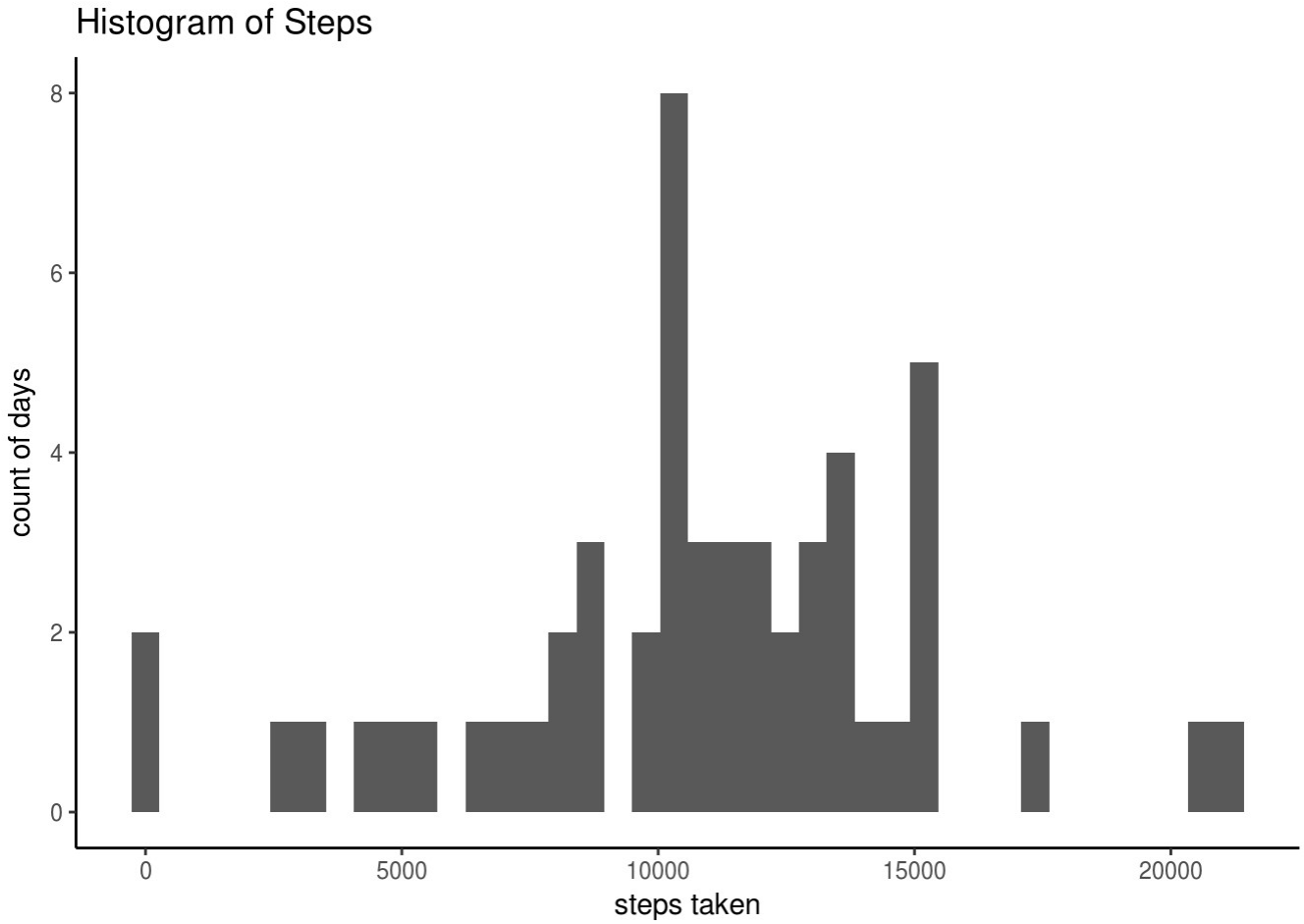
There are some intervals which have no step values. These NAs are removed from d and the result stored in clnd (clean d).

```
clnd <- d[is.na(d$steps)==FALSE,]
```

What is mean total number of steps taken per day?

1. Calculate the total number of steps taken per day

```
hist <- dt %>% group_by(date) %>% summarize(sum_steps=sum(steps))
qplot(x=hist$sum_steps,bins=40) +
  theme_classic() +
  labs(title='Histogram of Steps',y='count of days',x='steps taken')
```



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

The days with missing information are removed. Also to calculate the median 0 measurements are removed. The median and mean is as follows.

```
mean_total <- mean(dt[dt$steps>0,]$steps, na.rm=TRUE)
median_total <- median(dt[dt$steps>0,]$steps,na.rm=TRUE)

mean_total
```

```
## [1] 134.2607
```

```
median_total
```

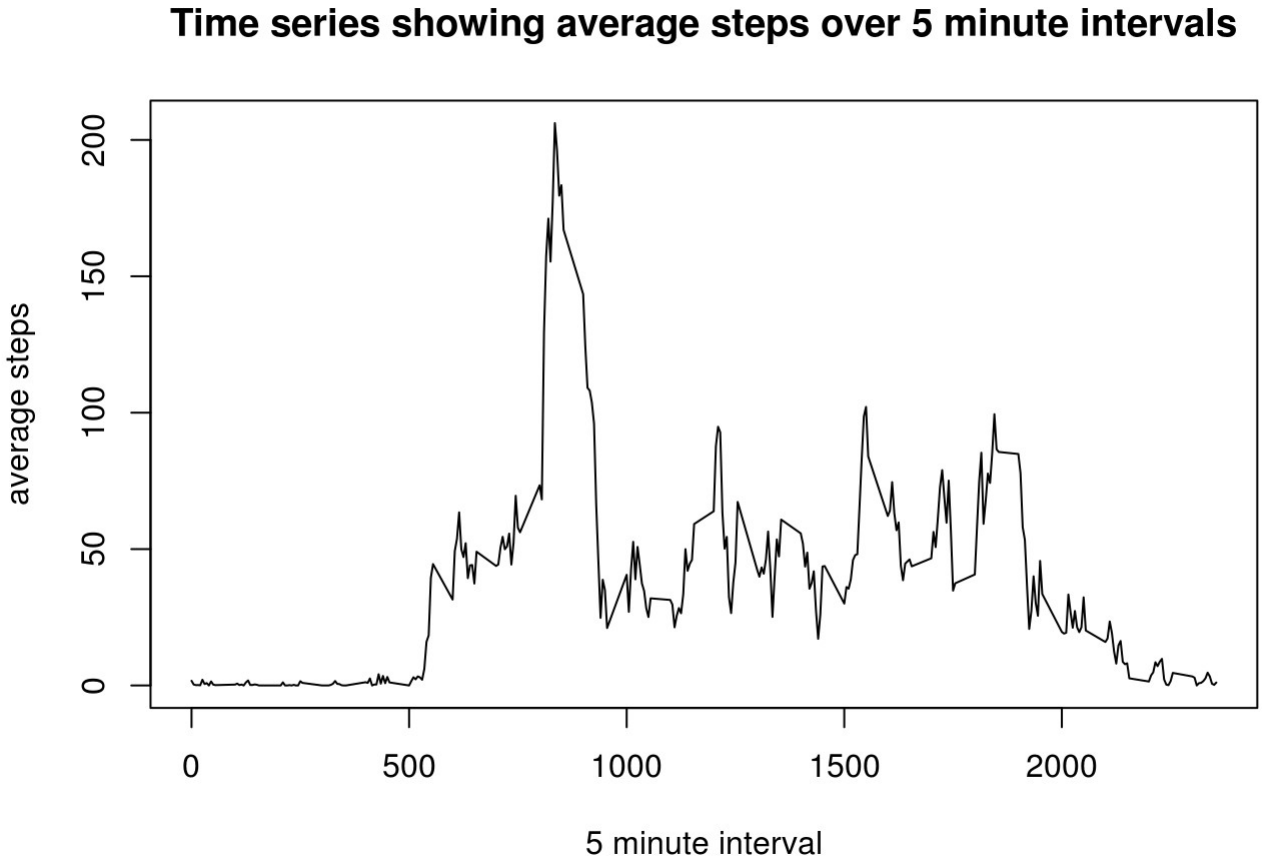
```
## [1] 56
```

What is the average daily activity pattern?

1. The average is calculated and a time series plot generated.

The average is calculated and a time series plot displayed based on the average.

```
average <- dt %>% group_by(interval) %>% summarize(avg_steps=mean(steps, na.rm=TRUE))

plot(x=average$interval, y=average$avg_steps,type='l',xlab = "5 minute interval", yla
b = "average steps",main="Time series showing average steps over 5 minute intervals")
```



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

The 5-minute interval with the highest average across all days is:

```
average[average$avg_steps==max(average$avg_steps),]$interval
```

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

Imputing missing values

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

The number of missing values follows.

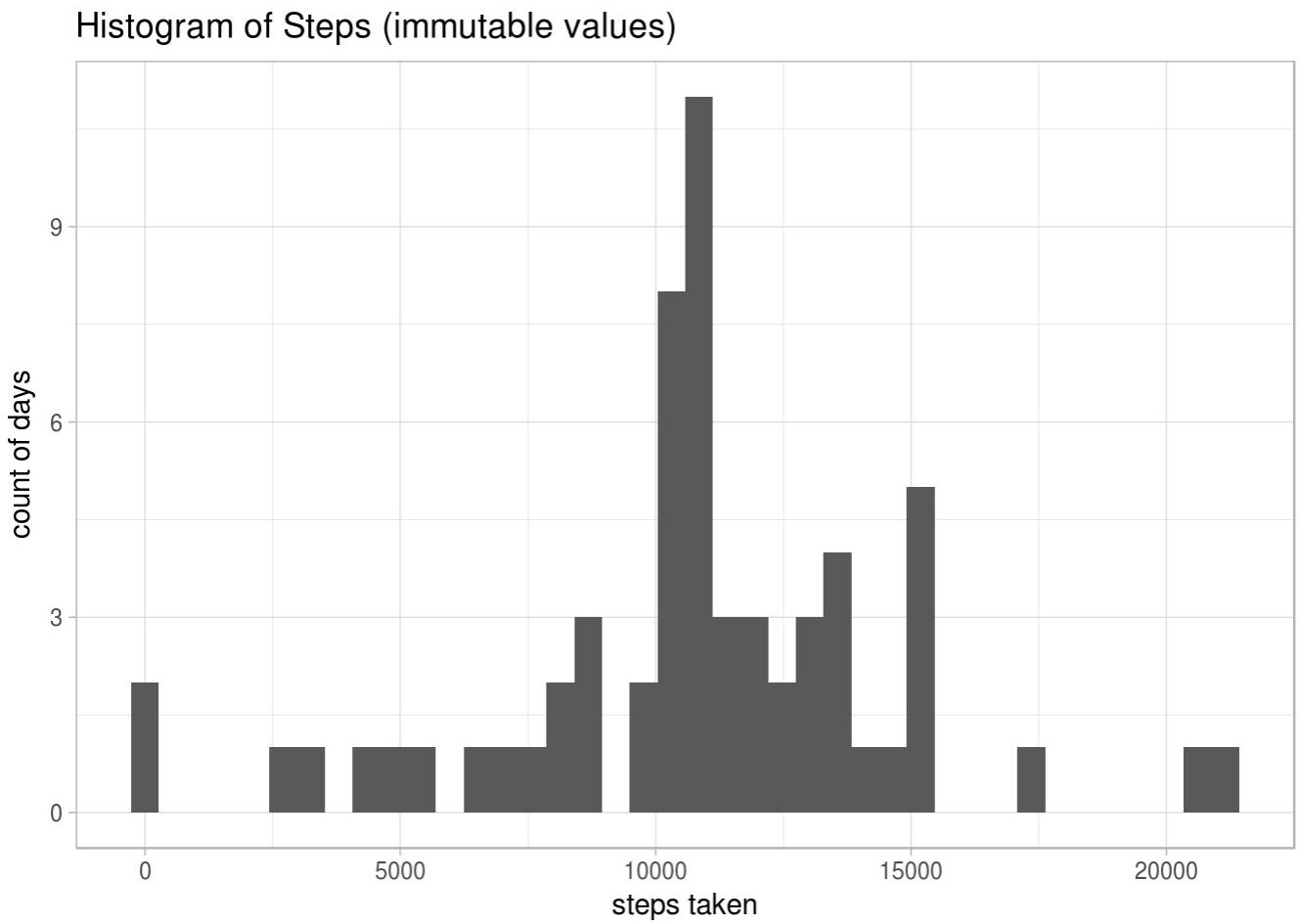
```
count(d[is.na(d$steps)==TRUE,])$n
```

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

2. Devise a strategy for filling in all of the missing values in the dataset.

The missing values are immuted by applying the average for the interval (across all days) to replace the missing values. There are some cleanup of variables. The immuted data set is merged with the original data set to create a new data set without missing values. A histogram is then generated. The total is used to calculate the mean and average. The median and mean is then displayed.

```
clnd %>% group_by(interval) %>% summarise(new_mean_steps = mean(steps)) -> avg_interv
al
d[is.na(d$steps)==TRUE,] -> only_missing
left_join(only_missing, avg_interval, by = 'interval') -> non_missing
non_missing[,c("steps")] <- NULL
rm(only_missing)
left_join(d, non_missing, by = c('date','interval')) -> d_non_missing
d_non_missing %>% group_by(date, interval) %>% summarise(immutable_step = max(steps,
new_mean_steps, na.rm = TRUE)) -> d_non_missing
rm(non_missing)

mean_total <- mean(d_non_missing$immutable_step, na.rm=TRUE)
median_total <- median(d_non_missing[d_non_missing$immutable_step>0,]$immutable_step,
na.rm=TRUE)

hist <- d_non_missing %>% group_by(date) %>% summarize(sum_steps=sum(immutable_step))
qplot(x=hist$sum_steps,bins=40) +
  theme_light() +
  labs(title='Histogram of Steps (immutable values)',y='count of days',x='steps taken
')
```




```
mean_total
```

```
## [1] 37.3826
```

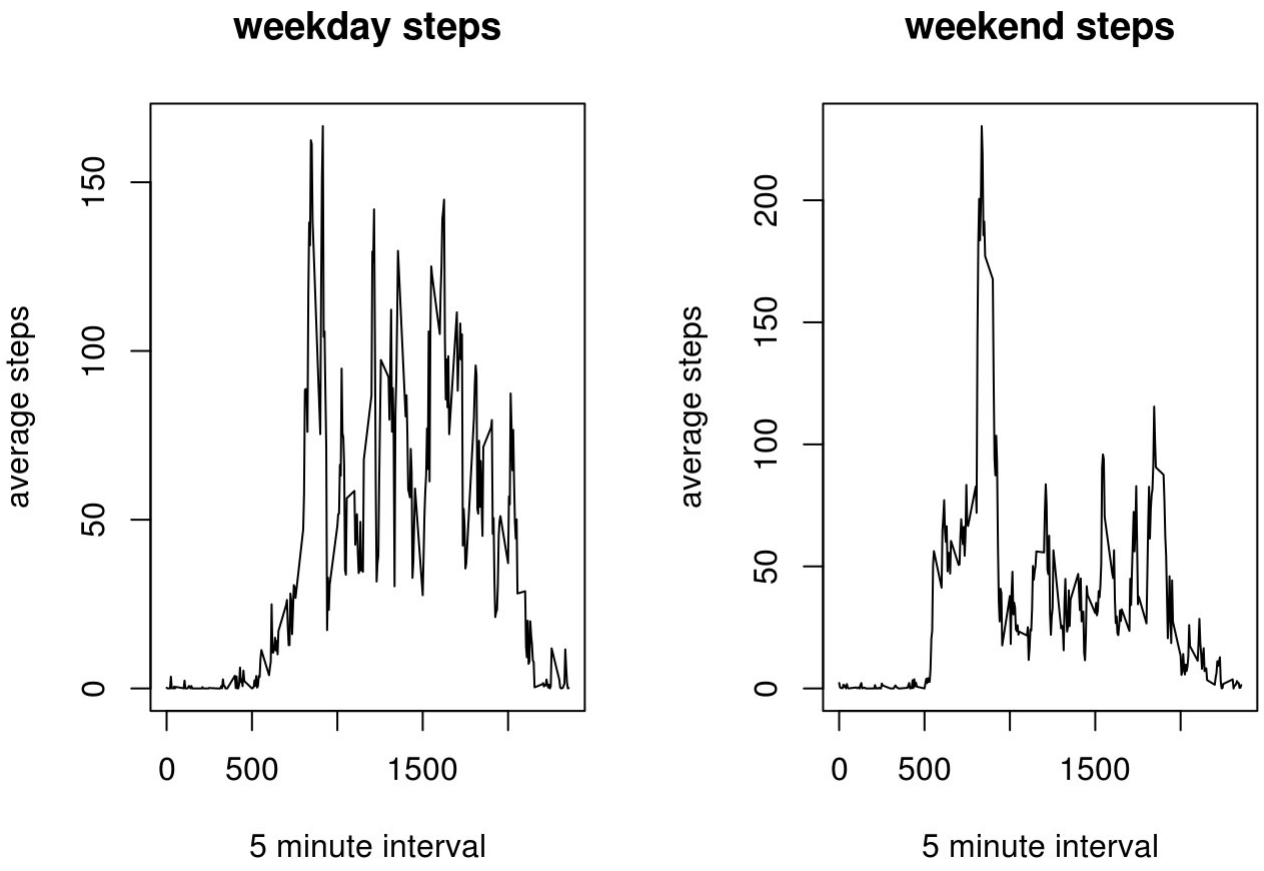
```
median_total
```

```
## [1] 45.33962
```

Are there differences in activity patterns between weekdays and weekends?

The first step is to determine which date the date corresponds to. Then determine if this date is a weekday. Plot as then generated for comparison.

```
d_non_missing[,'wd'] <- weekdays(as.Date(d_non_missing$date))
d_non_missing[,'wd_or_we'] <- (d_non_missing$wd == 'Sunday' | d_non_missing$wd == 'Sa
turday')

wd <- d_non_missing[d_non_missing$wd_or_we==TRUE,]
wd <- wd %>% group_by(interval) %>% summarize(avg_steps=mean(immutable_step, na.rm=TR
UE))

par(mfrow=c(1,2))

plot(x=wd$interval, y=wd$avg_steps,type='l',xlab = "5 minute interval", ylab = "avera
ge steps",main="weekday steps")

wd <- d_non_missing[d_non_missing$wd_or_we==FALSE,]
wd <- wd %>% group_by(interval) %>% summarize(avg_steps=mean(immutable_step, na.rm=TR
UE))

plot(wd$interval, wd$avg_steps,type='l',xlab = "5 minute interval", ylab = "average s
teps",main="weekend steps")
```



There are some differences with steps on the weekend and the steps on week days. The most active intervals are toward the begining of the day on weekends. On weekdays there is a gradual increase in steps which start a bit earlier while on weekends there is a sudden increase. Maybe the subject sleeps longer then does chores as soon as rising on weekends. The end of the assignment–I wonder, can the SQFT of the house be

determined by steps?