PA1_template

Nitin

July 4, 2019

```
knitr::opts_chunk$set(echo = 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(ggplot2)
```

1) Loading and preprocessing the data

1. Load the data (i.e. read.csv())

2. Process/transform the data (if necessary) into a format suitable for your analysis

```
data <- read.csv('activity.csv')
data$date <- as.Date(data$date)
```

```
## Warning in strptime(xx, f <- "%Y-%m-%d", tz = "GMT"): unable to identify current timezone
'C':
## please set environment variable 'TZ'
```

```
str(data)
```

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

```
head(str,5)
```

```
##
## 1 function (object, ...)
## 2 UseMethod("str")
```

2) What is mean total number of steps taken per day?

For this part of the assignment, you can ignore the missing values in the dataset.

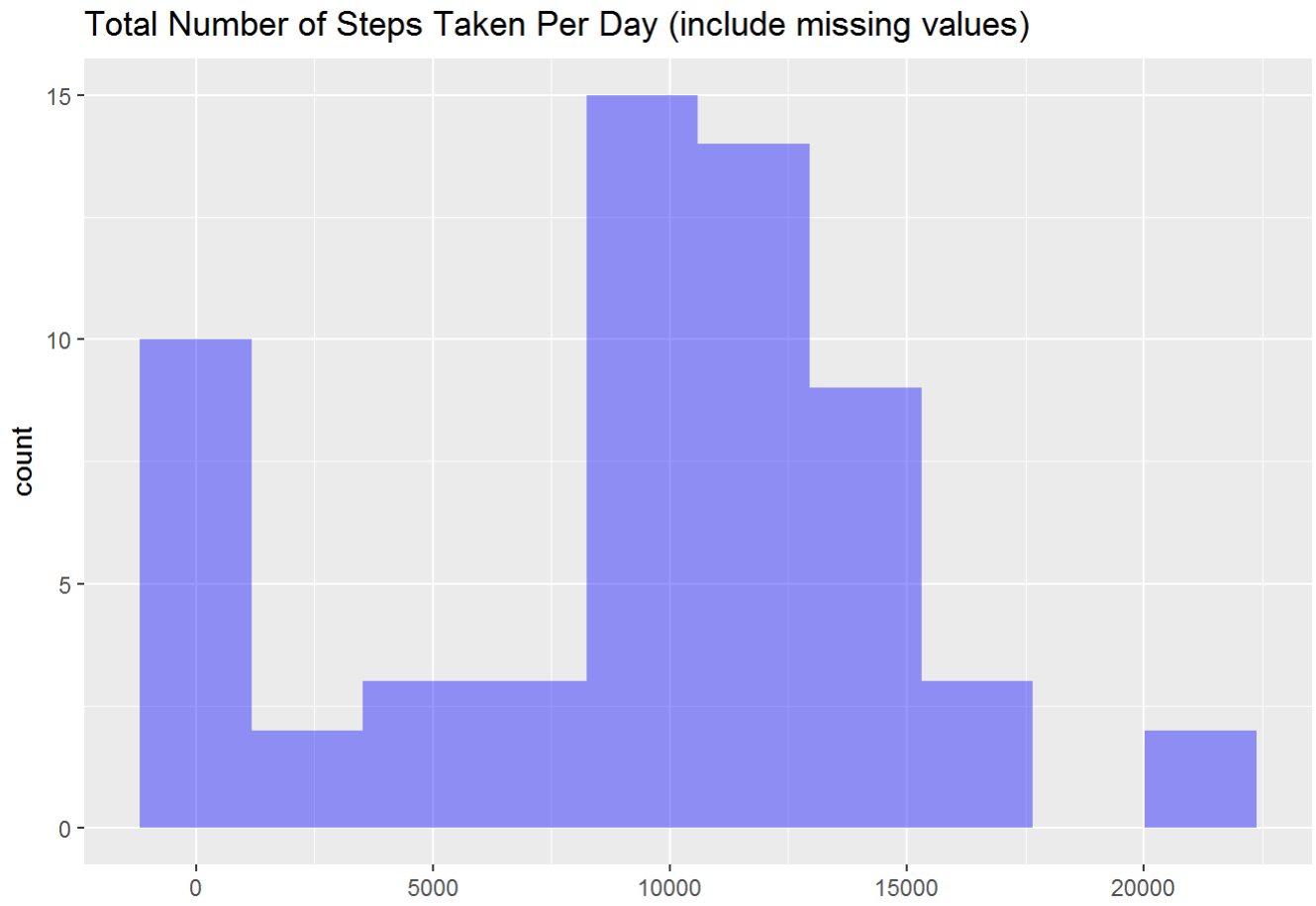
1. Calculate the total number of steps taken per day

```
datewise_mean <- data %>% group_by(date) %>% summarise(mean = mean(steps, na.rm = TRUE) )
```

2.Make a histogram of the total number of steps taken each day

```
  #agg <- aggregate(steps~date, data, sum)
  agg <- aggregate(data$steps, by = list(data$date), FUN = sum, na.rm=TRUE)
  mean.steps <- round(mean(agg$x, na.rm=TRUE))
  median.steps <- round(median(agg$x, na.rm=TRUE))

  ggplot(agg, aes(x = x)) +
  geom_histogram(bins = 10, fill = rgb(0, 0, 1, .4)) +
  xlab("") +
  ggtitle("Total Number of Steps Taken Per Day (include missing values)")
```



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

```
  paste("The Mean number of steps is",mean.steps,"and the Median is",median.steps)
```

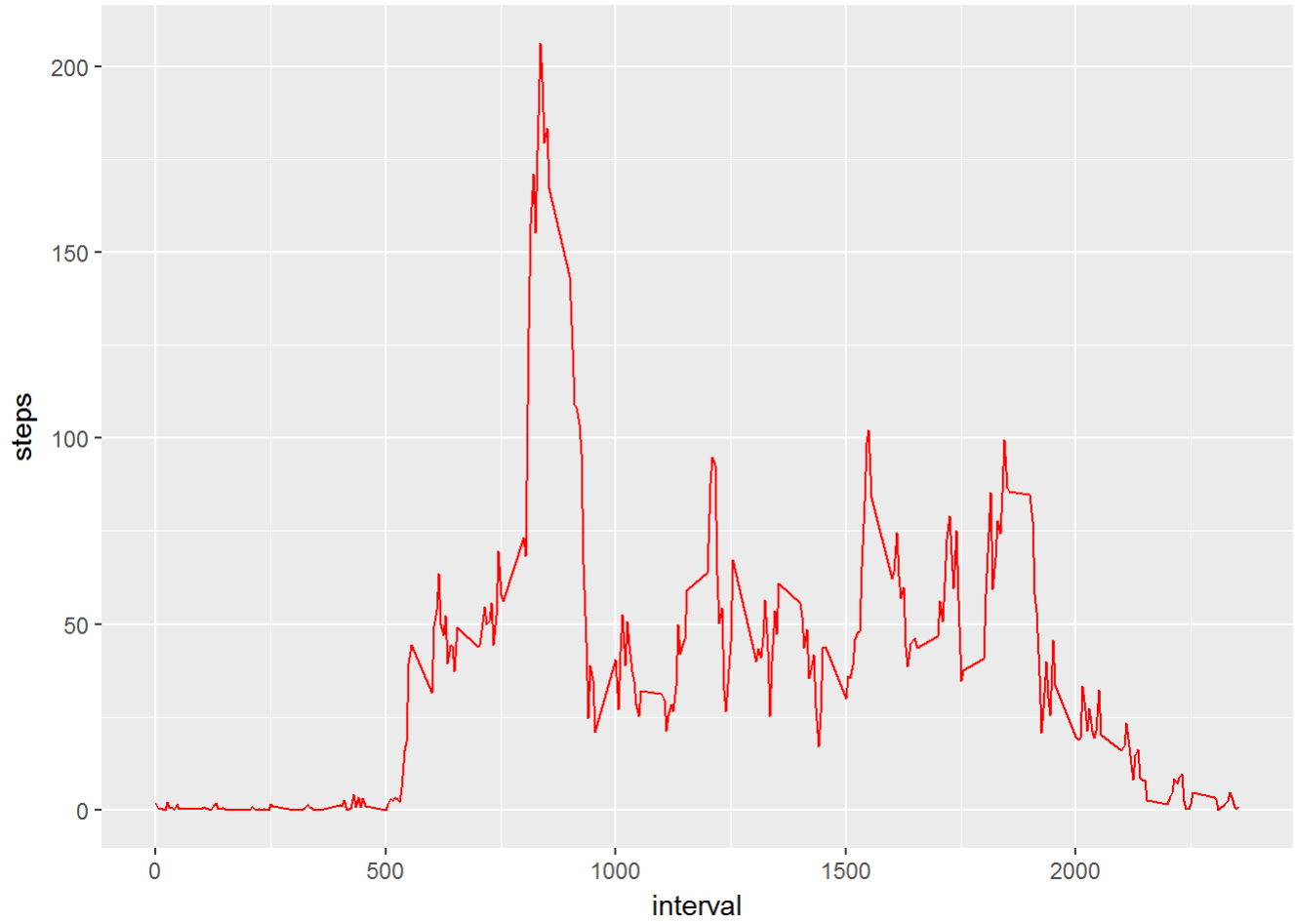
```
## [1] "The Mean number of steps is 9354 and the Median is 10395"
```

3) What is the average daily activity pattern?

1. 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)

```
agg_interval <- aggregate(x = list(steps = data$steps), by = list(interval=data$interval), FUN =
mean, na.rm=TRUE)

ggplot(agg_interval, aes(x = interval, y = steps)) +
  geom_line(color = "red")
```



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

```
max <- max(agg_interval$steps)
```

4) 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)

```
    paste("The number of missing values in the dataset is", sum(is.na(data$steps)) )
```

```
## [1] "The number of missing values in the dataset is 2304"
```

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

Filling all missing values by the mean for that 5-minute interval

Strategy :- Fill with Mean value Create a new dataset that is equal to the original dataset but with the missing data filled in.

```
interval_avg <- data %>%
  group_by(interval) %>%
  summarise(avg_steps = mean(steps, na.rm = TRUE))
```

3. Create a new dataset that is equal to the original dataset but with the missing data filled in.

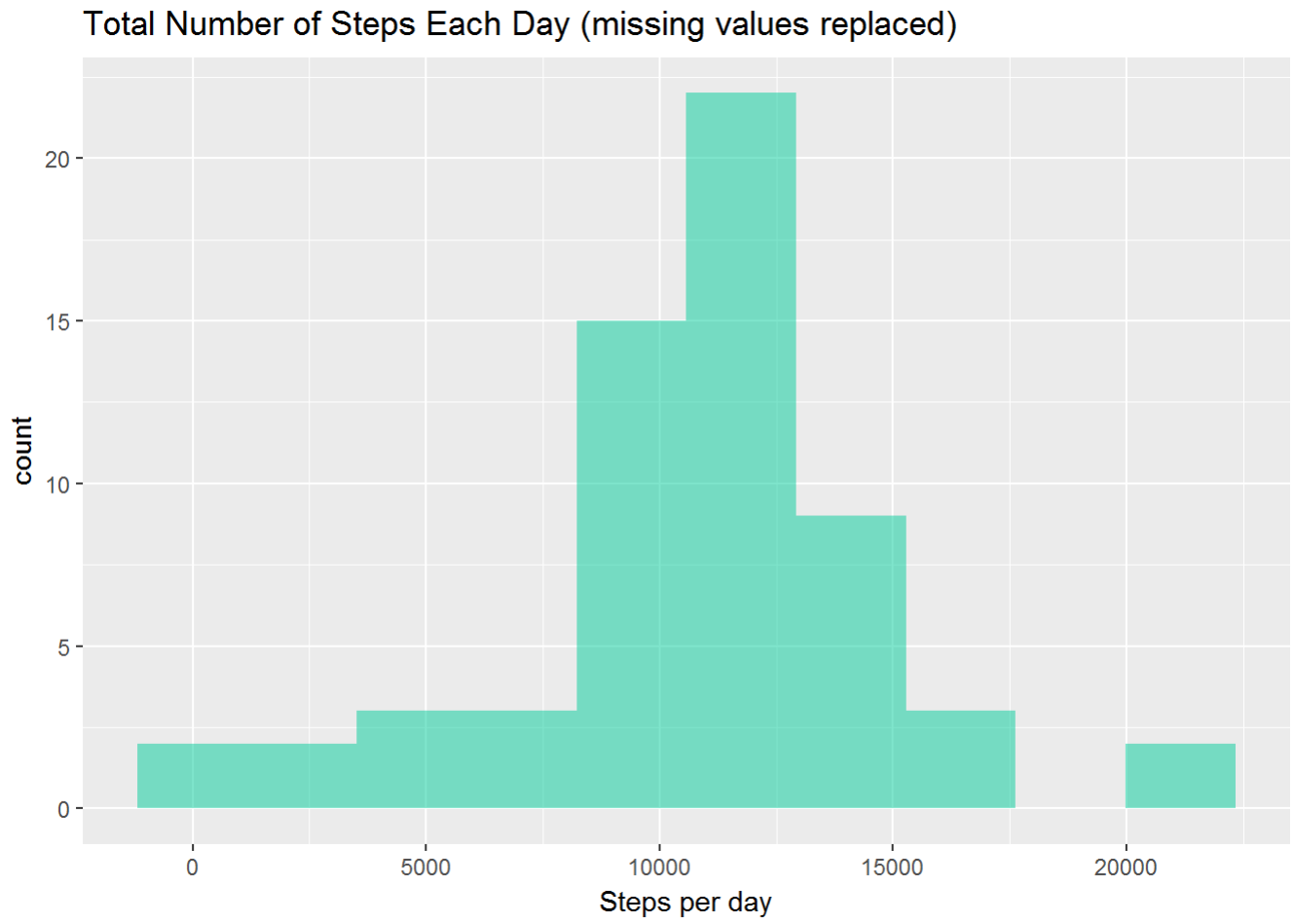
```
df_imputed <- data
for (i in 1:nrow(df_imputed)) {
  if (is.na(df_imputed[i, "steps"]) == TRUE) {
    data_interval <- df_imputed[i, "interval"]
    imputed_value <- interval_avg[interval_avg$interval == data_interval, "avg_steps"]
    df_imputed[i, "steps"] <- imputed_value
  } else {
    df_imputed[i, "steps"] <- df_imputed[i, "steps"]
  }
}
```

4. 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?

```
summarized_df <- df_imputed %>%
  group_by(date) %>%
  summarise(ttl_steps = sum(steps, na.rm = TRUE))

ggplot(summarized_df, aes(x = ttl_steps)) +
  geom_histogram(bins = 10, fill = rgb(0, .8, 0.6, 0.5)) +
  xlab("Steps per day") +
  ggtitle("Total Number of Steps Each Day (missing values replaced)")
```




```
mean_imputed_steps <- mean(summarized_df$ttl_steps)
median_imputed_steps <- median(summarized_df$ttl_steps)

 paste("The Mean number of steps is",mean_imputed_steps,"and the Median is",median_imputed_step
s)
```

```
## [1] "The Mean number of steps is 10766.1886792453 and the Median is 10766.1886792453"
```

5) Are there differences in activity patterns between weekdays and weekends?

For this part the weekdays() function may be of some help here. Use the dataset with the filled-in missing values for this part.

1. 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.

```
data$weekday_indicator <- ifelse(weekdays(data$date) %in% c("Saturday", "Sunday"), "weekend", "w
eekday")
```

2. 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).

```
summarized_df <- data %>%
  group_by(interval, weekday_indicator) %>%
  summarise(avg_steps = mean(steps, na.rm = TRUE))

ggplot(summarized_df, aes(x = interval, y = avg_steps)) +
  geom_line(color = "red") +
  facet_wrap(~weekday_indicator)
```