Analysis of Activity Monitoring Data

Daniel Aklilu

2025-07-04

##Introduction (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.)

#Loading the Data

```
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)
```

```
fileURL <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip"
```

```
download.file(fileURL, destfile = "repdata_data_activity.zip")
```

```
getwd()
```

```
unzip(zipfile = "./repdata_data_activity.zip", exdir = "./Project-7")
```

```
list.files("./Project-7")
```

```
file.exists("./Project-7/activity.csv")
```

```
DataSet <-read.csv("./Project-7/activity.csv")
```

```
print(DataSet)
```

Cleaning the data

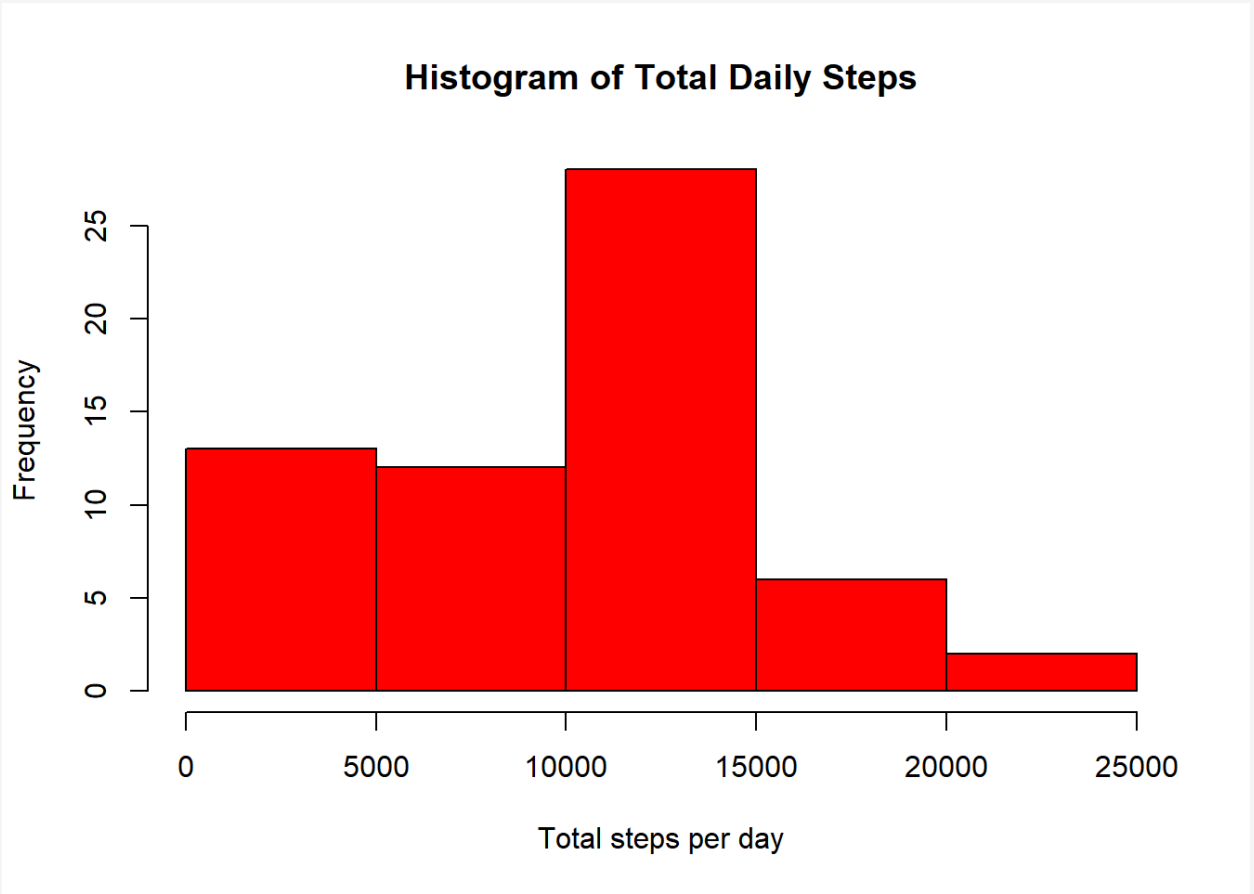
(Summarize the total number of steps taken per day.(Ignoring NA values for now))

```
steps_pd <- DataSet %>%
          group_by(date) %>%
          summarise(Total_Steps = sum(steps, na.rm = TRUE))
         print(steps_pd)
```

Histogram plot

(Use this to visualize the distribution of daily step counts.)

```
Plot1 <- hist(steps_pd$Total_Steps ,xlab = "Total steps per day", ylab = "Frequency", main = "Histogram of Total Daily Steps", col= "red")
```




```
summary(steps_pd$Total_Steps)
```

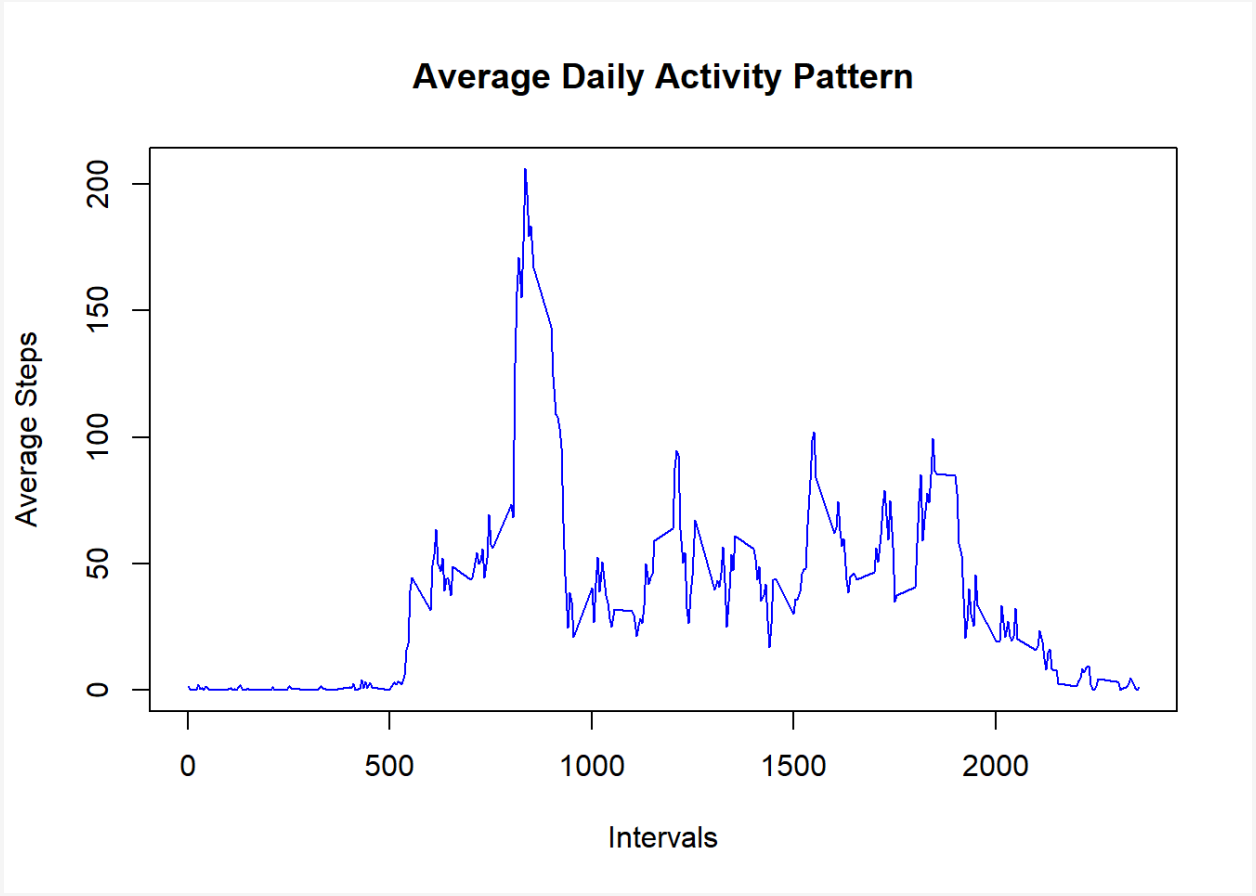
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    6778   10395    9354   12811   21194
```

Group the data by interval

```
Avg_Step <- DataSet %>%
    group_by(interval) %>%
    summarise(Average_Steps = mean(steps, na.rm = TRUE))
   print(Avg_Step)
```

#Make a Time series Plot

```
plot(x = Avg_Step$interval, y = Avg_Step$Average_Steps, type = "l",
 xlab = "Intervals", ylab = "Average Steps", main = "Average Daily Activity Pattern", col   = "blue")
```



#Find which one is the maximum value in interval

```
which.max(Avg_Step$Average_Steps)
```

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

```
Avg_Step$Average_Steps[104]
```

```
## [1] 206.1698
```

#Impute missing values

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

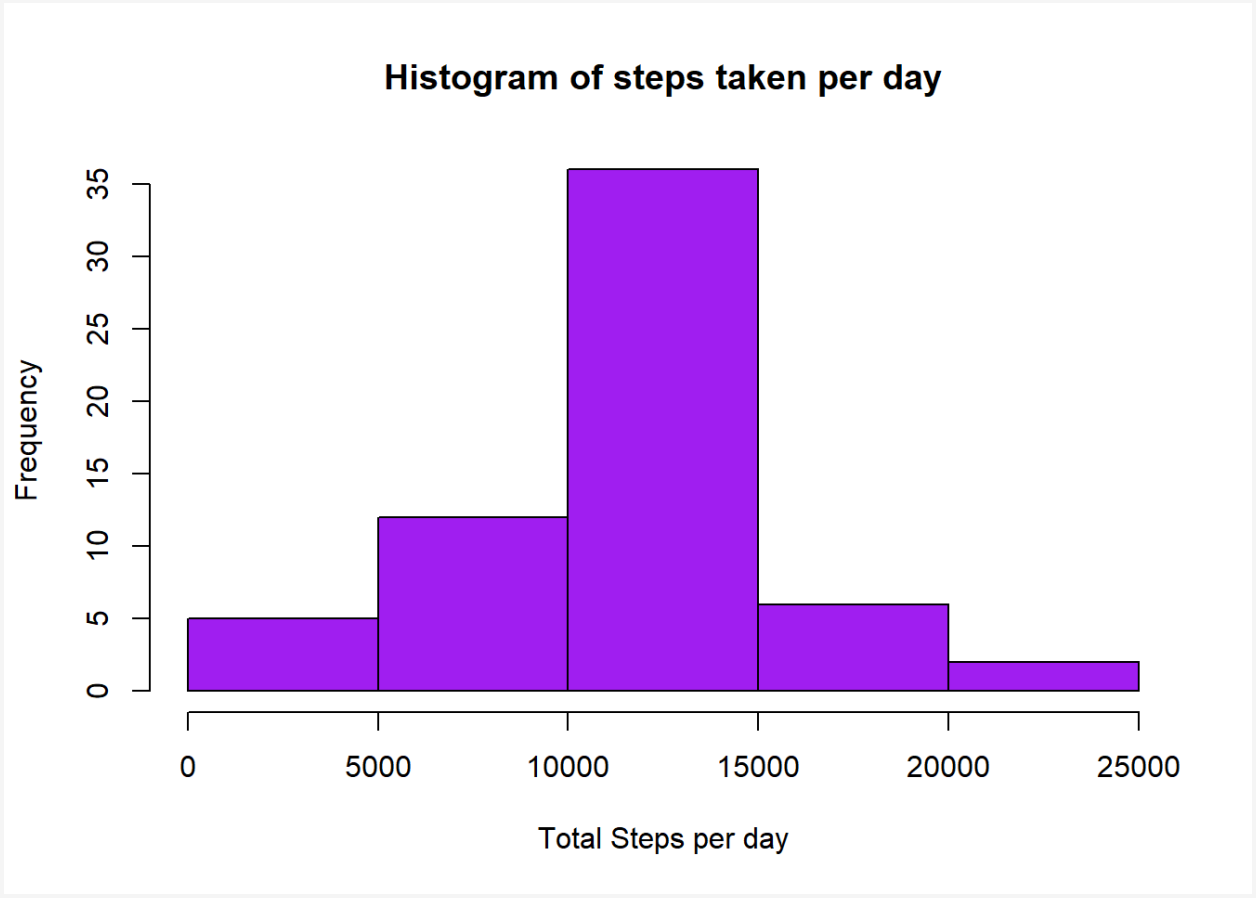
```
join_values <- left_join(DataSet,Avg_Step, by = "interval")
join_values <- join_values %>%
mutate(steps_filled = ifelse(is.na(steps), Average_Steps, steps))

print(join_values)
```

#Finding the total of number of steps per day.

```
 Steps <- join_values %>%
          group_by(date)%>%
          summarise(Total_steps = sum(steps_filled))

         print(Steps)
```

#Ploting a Histogram for the number of steps taken per day. (Calculate and print the mean and median number of total steps per day.)

```
      hist(Steps$Total_steps, xlab = "Total Steps per day", ylab = "Frequency", main = "Histogram of steps taken per day", col = "purple")
```




```
summary(Steps$Total_steps)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##     41    9819   10766   10766   12811   21194
```

Impact of imputing missing values

Imputing the missing values with the mean for each 5-minute inteval slightly changed the overalll distribution of total daily steps. The mean remained nearly the same, while the median showed a small shift. This suggests that the missing data were relatively evenly distributed and the imputation method preserved the original data characteristics.

#Label which is weekday and which is weekend

```
join_values$date <- as.Date(join_values$date)

Weekdays <- weekdays(join_values$date)

join_values <- join_values %>%
mutate(day_type = ifelse(weekdays(join_values$date) %in% c("Saturday",
"Sunday"),"weekends","weekdays"))

invisible(print(join_values))

table(join_values$day_type)
```

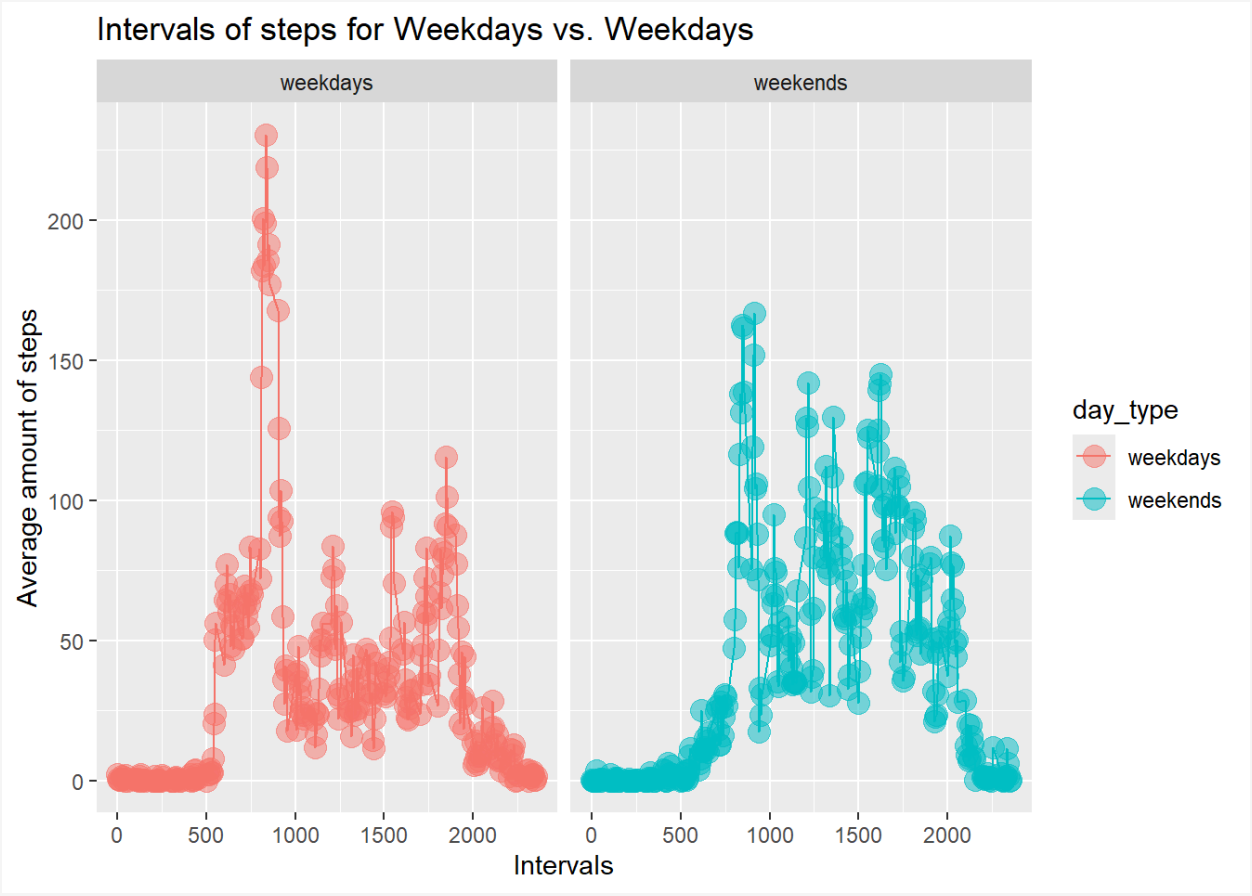
#Summaries the mean of steps for weekends and weekdays

```
interval_weekdays <- join_values%>%
        filter(join_values$day_type == "weekdays")%>%
        group_by(interval)%>%
       summarise(mean_steps=
        mean(steps_filled)) %>%
        mutate(day_type = "weekdays")

print(interval_weekdays)

interval_weekends <- join_values%>%
        filter(join_values$day_type == "weekends")%>%
        group_by(interval)%>%
        summarise(mean_steps=
        mean(steps_filled)) %>%
        mutate(day_type = "weekends")

print(interval_weekends)
```

#Creating a ggplot for weekends and weekdays

```
combined_days <- rbind(interval_weekdays,interval_weekends)
print(combined_days)
```

```
ggplot(combined_days, aes(interval, mean_steps, color = day_type)) + geom_line() + geom_point(size = 4, alpha =0.5) + facet_wrap(~day_type) + labs( x = "Intervals", y = "Average amount of steps", title = "Intervals of steps for Weekdays vs. Weekdays")
```



Weekdays vs Weekend step Activities (On weekdays, there is a sharp spike in activity around 8:00 AM, which likely corresponds to morning routines commuting, walking to work/school.After it tends to drop off, with a few smaller peaks later in the day. This suggests that steps taken are more concentrated around specific times, probably due to work or school schedules. On The Weekends how ever the step activity is evenly distributed throughout the day. There is no sharp morning peak, indicating that people likely wake up later or have more flexible hobbies, a moderate amount of steps are taken later in the day, suggesting more consistent movement, possibly from leisure activities or errands.)