Reproducible Research Project 1

Carla F

27/12/2020

Packages used for this project:

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)

Import of the dataset

ds <- read.csv("activity.csv")  
head(ds)

## steps date interval  
## 1 NA 2012-10-01 0  
## 2 NA 2012-10-01 5  
## 3 NA 2012-10-01 10  
## 4 NA 2012-10-01 15  
## 5 NA 2012-10-01 20  
## 6 NA 2012-10-01 25

What is mean total number of steps taken per day?

1. Calculate the total number of steps taken per day

totalstep <- ds %>% select(steps,date) %>%  
 group\_by(date) %>%  
 summarise(steps = mean(steps, na.rm = TRUE))

## `summarise()` ungrouping output (override with `.groups` argument)

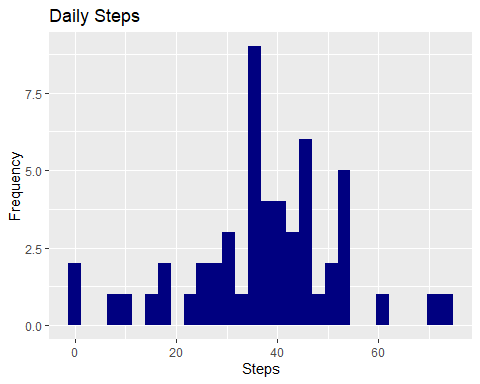
totalstep$steps <- as.numeric(totalstep$steps)  
head(totalstep)

## # A tibble: 6 x 2  
## date steps  
## <chr> <dbl>  
## 1 2012-10-01 NaN   
## 2 2012-10-02 0.438  
## 3 2012-10-03 39.4   
## 4 2012-10-04 42.1   
## 5 2012-10-05 46.2   
## 6 2012-10-06 53.5

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

ggplot(totalstep,aes(x=steps)) +   
 geom\_histogram(fill="navy", bins=30)+  
 labs(title = "Daily Steps", x="Steps", y="Frequency")

## Warning: Removed 8 rows containing non-finite values (stat\_bin).



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

totalstep1 <- ds %>% select(steps,date) %>%  
 group\_by(date) %>%  
 summarise(meansteps = mean(steps, na.rm = TRUE),  
 mediansteps = median(steps, na.rm = TRUE))

## `summarise()` ungrouping output (override with `.groups` argument)

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)

intervalstep <- ds %>% select(steps,interval) %>%  
 group\_by(interval) %>%  
 summarise(steps = mean(steps, na.rm = TRUE))

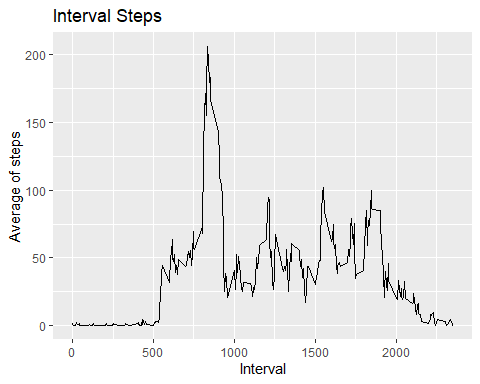
## `summarise()` ungrouping output (override with `.groups` argument)

intervalstep$steps <- as.numeric(intervalstep$steps)  
head(intervalstep)

## # A tibble: 6 x 2  
## interval steps  
## <int> <dbl>  
## 1 0 1.72   
## 2 5 0.340   
## 3 10 0.132   
## 4 15 0.151   
## 5 20 0.0755  
## 6 25 2.09

ggplot(intervalstep,aes(x=interval,y=steps))+  
 geom\_line(fill="dark green")+  
 labs(title = "Interval Steps", x="Interval", y="Average of steps")

## Warning: Ignoring unknown parameters: fill



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

intervalstep %>% filter(steps == max(steps))

## # A tibble: 1 x 2  
## interval steps  
## <int> <dbl>  
## 1 835 206.

Imputing missing values

1. Calculate and report the total number of missing values in the dataset

nas <- is.na(ds$steps)  
sum(nas)

## [1] 2304

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

mean <- ds %>% summarise(steps = mean(steps, na.rm = TRUE))  
Test <- unique(ds$interval)  
ds\_notNA <- ds  
for (i in 1:length(Test)) {  
 ds\_notNA$steps[is.na(ds\_notNA$steps) & ds\_notNA$interval == Test[i]] <- mean  
}

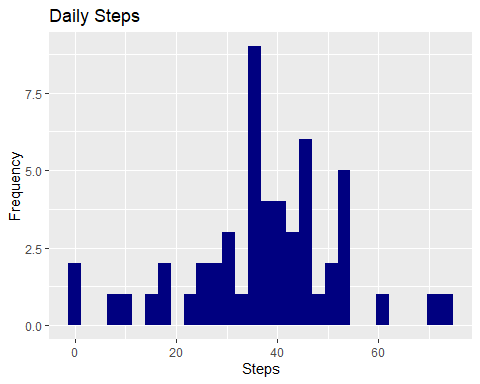
1. Create a new dataset that is equal to the original dataset but with the missing data filled in

newds <- for (i in 1:length(Test)) {  
 ds\_notNA$steps[is.na(ds\_notNA$steps) & ds\_notNA$interval == Test[i]] <- mean[i,2]  
}

1. 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?

ggplot(totalstep,aes(x=steps)) +   
 geom\_histogram(fill="navy", bins=30)+  
 labs(title = "Daily Steps", x="Steps", y="Frequency")

## Warning: Removed 8 rows containing non-finite values (stat\_bin).



Are there differences in activity patterns between weekdays and weekends?

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

ds <- read.csv("activity.csv")  
ds$date <- as.Date(ds$date)  
weekdays1 <- c('Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday')  
  
ds$wDay <- factor((weekdays(ds$date) %in% weekdays1), levels=c(FALSE, TRUE), labels=c('weekend', 'weekday'))  
  
head(ds)

## steps date interval wDay  
## 1 NA 2012-10-01 0 weekday  
## 2 NA 2012-10-01 5 weekday  
## 3 NA 2012-10-01 10 weekday  
## 4 NA 2012-10-01 15 weekday  
## 5 NA 2012-10-01 20 weekday  
## 6 NA 2012-10-01 25 weekday

1. 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)

ggplot(ds, aes(interval, steps, col=wDay)) + geom\_line() + facet\_grid(~wDay)+  
 xlab("Intervals") + ylab("Frequency of steps")

## Warning: Removed 2 row(s) containing missing values (geom\_path).

