|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  | | --- | --- | --- | --- | | Title | Author | Date | Output | | Reproducible Research – Assessment 1 | Thafne Plastina Astro | 06/12/2022 | **html\_document**  **keep\_md** | |  |  |  |  | |  |  |  |

knitr::opts\_chunk$set(fig.path='Figs/')

**Introduction**

This project is part of the course Reproducible Researche on Coursera.

The assignment makes use of data collected from a personal activity monitoring device. This device collects data at 5 minutes interval throughout 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.

**Data**

The data for this assignment can be downloaded from the course web site:

* Dataset: [Activity monitoring data](https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2Factivity.zip) [52K]

The variables included in this dataset are:

* **steps:** Number of steps taking in a 5-minute interval (missing values are coded as NA)
* **date:** The date on which the measurement was taken in YYYY-MM-DD format
* **interval:** Identifier for the 5-minute interval in which measurement was taken

The dataset is stored in a comma-separated-value (CSV) file and there are a total of 17,568 observations in this dataset.

**Load libraries**

library(dplyr)

library(lattice)

**Load and preprocess the data**

#load data

data<-read.csv("activity.csv")

#format dates

data$date<-as.Date(data$date, format="%Y-%m-%d")

**What is mean total number of steps taken per day?**

**Make a histogram and report the mean and median of the total number of steps taken per day**

#group data by date and sum steps

grouped\_day<-group\_by(data, data$date)

day\_total<-summarise(grouped\_day, "sum"=sum(steps, na.rm=T))

#plot histogram

hist(day\_total$sum, breaks=35, main="Total number of steps taken per day", xlab="Total number of steps", col="gray")

abline(v=mean(day\_total$sum), col="red", lwd=2)

abline(v=median(day\_total$sum), col="blue", lwd=2)

legend("topright", c("mean: 9354", "median: 10395"), col=c("red", "blue"), lwd=2)

Gráfico

Descrição gerada automaticamente

## What is the average daily activity pattern?

#### Make a time series plot of the 5-minute interval and the average number of steps taken, averaged across all days. Which 5-minute interval, on average across all the days in the dataset, contains the maximum number of steps?

#group data by interval and calculate mean

grouped\_interval<-group\_by(data, data$interval)

interval\_avg<-summarise(grouped\_interval, "mean"=mean(steps, na.rm=T))

#plot graph

plot(interval\_avg, type="l", main="Averge number of steps taken per interval", xlab="Interval", ylab="Average number of steps")

points(x=interval\_avg[interval\_avg$mean==max(interval\_avg$mean),1], y=max(interval\_avg$mean), pch=19, col="red")

text(x=1300, y=190, labels="Maximun average number \nof steps (206 steps) \nat the 835th interval", cex = 0.8)

Gráfico, Histograma

Descrição gerada automaticamente

Across all the days in the dataset, the interval with highest average number of steps is the 835th, with 206 average steps.

## Imputing missing values

#### Calculate and report the total number of missing values in the dataset

sum(is.na(data$steps))

## [1] 2304

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

I will replace the NA's with the median number of steps for a given interval.

#create a data frame with the median number of steps for each interval

interval\_median<-summarise(grouped\_interval, "median"=median(steps, na.rm=T))

#merge it with original data frame by the interval

data\_replaced<-merge(data, interval\_median, by.x="interval", by.y="data$interval")

#replace NA's with median for that interval

data\_replaced$steps[is.na(data\_replaced$steps)]<-data\_replaced$median[is.na(data\_replaced$steps)]

#### 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.

#group new data by date and sum steps

grouped\_day\_replaced<-group\_by(data\_replaced, data\_replaced$date)

day\_total\_replaced<-summarise(grouped\_day\_replaced, "sum"=sum(steps))

#change parameters to plot both graphs together for comparasion

par(mfrow=c(1,2), mar=c(5,5,5,2))

#plot histogram with original data (with NA)

hist(day\_total$sum, breaks=35, main="Original", xlab="Total number of steps", col="gray")

abline(v=mean(day\_total$sum), col="red", lwd=2)

abline(v=median(day\_total$sum), col="blue", lwd=2)

legend("topright", c(paste("mean: ", round(mean(day\_total$sum)), 2), paste("median: ", round(median(day\_total$sum)), 2)), col=c("red", "blue"), lwd=2, cex = 0.7)

#plot histogram with NA's replaced

hist(day\_total\_replaced$sum, breaks=35, main="Replaced", xlab="Total number of steps", col="gray")

abline(v=mean(day\_total\_replaced$sum), col="red", lwd=2)

abline(v=median(day\_total\_replaced$sum), col="blue", lwd=2)

legend("topright", c(paste("mean: ", round(mean(day\_total\_replaced$sum)), 2), paste("median: ", round(median(day\_total\_replaced$sum)),

Gráfico, Histograma

Descrição gerada automaticamente

#### Do these values differ from the estimates from the first part of the assignment?

We can see that there are far less zeros, which led the mean to change slightly but the median remained the same. Depending on the strategy adopted to fill the missing data, the results might differ. Having missing data may introduce bias in the analysis, but filling it may not rid the analysis from bias.

#### What is the impact of imputing missing data on the estimates of the total daily number of steps?

summary(day\_total$sum)

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

## 0 6778 10395 9354 12811 21194

summary(day\_total\_replaced$sum)

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

## 41 6778 10395 9504 12811 21194

As we can see in the histograms, the values change slightly, but the five number summary remains pretty much the same.

## Weekdays and weekends

#### Are there differences in activity patterns between weekdays and weekends?

#create weekday variable

data\_replaced$weekday<-weekdays(data\_replaced$date)

#create two data frames, one for week days and another for weekends

week\_dates<-data\_replaced[which(data\_replaced$weekday %in% c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday")),]

weekend\_dates<-data\_replaced[which(data\_replaced$weekday %in% c("Saturday", "Sunday")),]

#plot graphs

par(mfrow=c(1,2))

plot(steps~interval, data=week\_dates, type="l", main="Week days")

plot(steps~interval, data=weekend\_dates, type="l", main="Weekends")

Gráfico, Histograma

Descrição gerada automaticamente

It seem like there is a difference in peak hours between weekdays and weekends. In weekdays, there is a higher density of number of steps in the morning, whereas in the weekend the highest density of number of steps is around the middle of the afternoon.