test

student

2023-08-10

```
# Set default options for code chunks
knitr::opts_chunk$set(
  echo = TRUE,          # Display R code and its output
  comment=NA,           # Suppress code comments in output
  warning = FALSE,      # Suppress warning messages
  fig.align='center',   # Align figures in the center
  eval = TRUE           # Evaluate R code
)
```

```
knitr::opts_chunk$set(
  echo = TRUE,
  comment=NA,
  warning = FALSE,
  fig.align='center',
  eval = TRUE
)
```

```
# Import libraries
library(tidyverse) # includes ggplot2
library(skimr) #  provides a compact and informative summary of your dataframe or dataset
library(lubridate)
library(janitor) # set of utility functions for data cleaning and data frame tidying tasks
library(RColorBrewer) # Color palettes for data visualization
library(ggcorrplot) # Visualize correlation matrices using ggplot2
library(scales) # formatting and transforming data for visualizations
```

Let us clean:

- Change column names to lower case because R is case sensitive.
- Change "Id" from double to a character because the number represents a category.
- Change "ActivityDate" from char to date.

```
weight_logs <-
  read_csv("original_data/weightLogInfo_merged.csv",
    trim_ws = TRUE,
    show_col_types = FALSE
  )
```

```
glimpse(weight_logs)
```

```
Rows: 67
Columns: 8
$ Id             <dbl> 1503960366, 1503960366, 1927972279, 2873212765, 2873212~
$ Date           <chr> "5/2/2016 11:59:59 PM", "5/3/2016 11:59:59 PM", "4/13/2~
```

```
$ WeightKg       <dbl> 52.6, 52.6, 133.5, 56.7, 57.3, 72.4, 72.3, 69.7, 70.3, ~
$ WeightPounds   <dbl> 115.9631, 115.9631, 294.3171, 125.0021, 126.3249, 159.6~
$ Fat            <dbl> 22, NA, NA, NA, NA, 25, NA, NA, NA, NA, NA, NA, NA, NA,~
$ BMI            <dbl> 22.65, 22.65, 47.54, 21.45, 21.69, 27.45, 27.38, 27.25,~
$ IsManualReport <lgl> TRUE, TRUE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, ~
$ LogId          <dbl> 1.462234e+12, 1.462320e+12, 1.460510e+12, 1.461283e+12,~
```

```
glimpse(weight_logs)
```

```
Rows: 67
Columns: 8
$ Id             <dbl> 1503960366, 1503960366, 1927972279, 2873212765, 2873212~
$ Date           <chr> "5/2/2016 11:59:59 PM", "5/3/2016 11:59:59 PM", "4/13/2~
$ WeightKg       <dbl> 52.6, 52.6, 133.5, 56.7, 57.3, 72.4, 72.3, 69.7, 70.3, ~
$ WeightPounds   <dbl> 115.9631, 115.9631, 294.3171, 125.0021, 126.3249, 159.6~
$ Fat            <dbl> 22, NA, NA, NA, NA, 25, NA, NA, NA, NA, NA, NA, NA, NA,~
$ BMI            <dbl> 22.65, 22.65, 47.54, 21.45, 21.69, 27.45, 27.38, 27.25,~
$ IsManualReport <lgl> TRUE, TRUE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, ~
$ LogId          <dbl> 1.462234e+12, 1.462320e+12, 1.460510e+12, 1.461283e+12,~
```

Observations:

- Many variables show a right-skewed distribution: a larger number of data values are located on the left side of the curve.
- The variables total_steps, total_distance, tracker_distance have a similar distribution. We can explore their correlations later.
- Since the distributions are not normal. The median is a better indicator of central tendency for the numerical variables in these dataset.
- **The variable logged_activities_distance and sedentary_active_distance might not provide useful information since most of the data points are zero. It seems that the users are not logging the distance frequently**
- The following variables seem related. We will explore them further in the bivariate analysis section:
  - sedentary_minutes; sedentary_active_distance
  - lightly_active_minutes; light_active_distance
  - fairly_active_minutes; moderately_active_distance
  - very_active_minutes; very_active_distance
- The variables calories and sedentary_minutes exhibit a multimodal distribution, indicating the presence of subpopulations within the data. In this dataset, gender could be a potential variable that would result in a bimodal distribution when examining histograms of calories and sedentary minutes. Unfortunately, the gender of the users is not provided, limiting our ability to confirm this hypothesis.