Example Starter R Script

Introduction and background

This is meant to be a sample starter script if you choose to use R for this case study. This is not comprehensive of everything you'll do in the case study, but should be used as a starting point if it is helpful for you.

Upload your CSV files to R

Remember to upload your CSV files to your project from the relevant data source: https://www.kaggle.com/arashnic/fitbit

Remember, there are many different CSV files in the dataset. We have uploaded two CSVs into the project, but you will likely want to use more than just these two CSV files.

Installing and loading common packages and libraries

You can always install and load packages along the way as you may discover you need different packages after you start your analysis. If you already have some of these packages installed and loaded, you can skip those ones - or you can choose to run those specific lines of code anyway. It may take a few moments to run.

```
# install.packages('tidyverse')
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.0     v forcats 0.5.1
```

```
## Warning: package 'forcats' was built under R version 4.1.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Loading your CSV files

Here we'll create a dataframe named 'daily_activity' and read in one of the CSV files from the dataset. Remember, you can name your dataframe something different, and you can also save your CSV file under a different name as well.

```
daily_activity <- read.csv("dailyActivity_merged.csv")
```

We'll create another dataframe for the sleep data.

```
sleep_day <- read.csv("sleepDay_merged.csv")
```

Exploring a few key tables

Take a look at the daily_activity data.

```
head(daily_activity)
```

```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366    4/12/2016      13162          8.50            8.50
## 2 1503960366    4/13/2016      10735          6.97            6.97
## 3 1503960366    4/14/2016      10460          6.74            6.74
## 4 1503960366    4/15/2016       9762          6.28            6.28
## 5 1503960366    4/16/2016      12669          8.16            8.16
## 6 1503960366    4/17/2016       9705          6.48            6.48
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0              1.88                     0.55
## 2                        0              1.57                     0.69
## 3                        0              2.44                     0.40
## 4                        0              2.14                     1.26
## 5                        0              2.71                     0.41
## 6                        0              3.19                     0.78
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                6.06                       0                25
## 2                4.71                       0                21
## 3                3.91                       0                30
## 4                2.83                       0                29
## 5                5.04                       0                36
## 6                2.51                       0                38
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  13                  328              728     1985
## 2                  19                  217              776     1797
## 3                  11                  181             1218     1776
## 4                  34                  209              726     1745
## 5                  10                  221              773     1863
## 6                  20                  164              539     1728
```

Identify all the columsn in the daily_activity data.

```
colnames(daily_activity)
```

```
##  [1] "Id"                       "ActivityDate"
##  [3] "TotalSteps"               "TotalDistance"
##  [5] "TrackerDistance"          "LoggedActivitiesDistance"
##  [7] "VeryActiveDistance"       "ModeratelyActiveDistance"
##  [9] "LightActiveDistance"      "SedentaryActiveDistance"
```

```
## [11] "VeryActiveMinutes"        "FairlyActiveMinutes"
## [13] "LightlyActiveMinutes"     "SedentaryMinutes"
## [15] "Calories"
```

Take a look at the sleep_day data.

```
head(sleep_day)
```

```
##           Id              SleepDay TotalSleepRecords TotalMinutesAsleep
## 1 1503960366 4/12/2016 12:00:00 AM                 1                327
## 2 1503960366 4/13/2016 12:00:00 AM                 2                384
## 3 1503960366 4/15/2016 12:00:00 AM                 1                412
## 4 1503960366 4/16/2016 12:00:00 AM                 2                340
## 5 1503960366 4/17/2016 12:00:00 AM                 1                700
## 6 1503960366 4/19/2016 12:00:00 AM                 1                304
##   TotalTimeInBed
## 1            346
## 2            407
## 3            442
## 4            367
## 5            712
## 6            320
```

Identify all the columsn in the daily_activity data.

```
colnames(sleep_day)
```

```
## [1] "Id"                 "SleepDay"           "TotalSleepRecords"
## [4] "TotalMinutesAsleep" "TotalTimeInBed"
```

Note that both datasets have the 'Id' field - this can be used to merge the datasets.

Understanding some summary statistics

How many unique participants are there in each dataframe? It looks like there may be more participants in the daily activity dataset than the sleep dataset.

```
n_distinct(daily_activity$Id)
```

```
## [1] 33
```

```
n_distinct(sleep_day$Id)
```

```
## [1] 24
```

How many observations are there in each dataframe?

```
nrow(daily_activity)
```

```
## [1] 940
```

```
nrow(sleep_day)
```

```
## [1] 413
```

What are some quick summary statistics we'd want to know about each data frame?

For the daily activity dataframe:

```
daily_activity %>%
  select(TotalSteps,
         TotalDistance,
         SedentaryMinutes) %>%
  summary()
```

```
##    TotalSteps    TotalDistance    SedentaryMinutes
##  Min.   :    0   Min.   : 0.000   Min.   :   0.0
##  1st Qu.: 3790   1st Qu.: 2.620   1st Qu.: 729.8
##  Median : 7406   Median : 5.245   Median :1057.5
##  Mean   : 7638   Mean   : 5.490   Mean   : 991.2
##  3rd Qu.:10727   3rd Qu.: 7.713   3rd Qu.:1229.5
##  Max.   :36019   Max.   :28.030   Max.   :1440.0
```

For the sleep dataframe:

```
sleep_day %>%
  select(TotalSleepRecords,
  TotalMinutesAsleep,
  TotalTimeInBed) %>%
  summary()
```

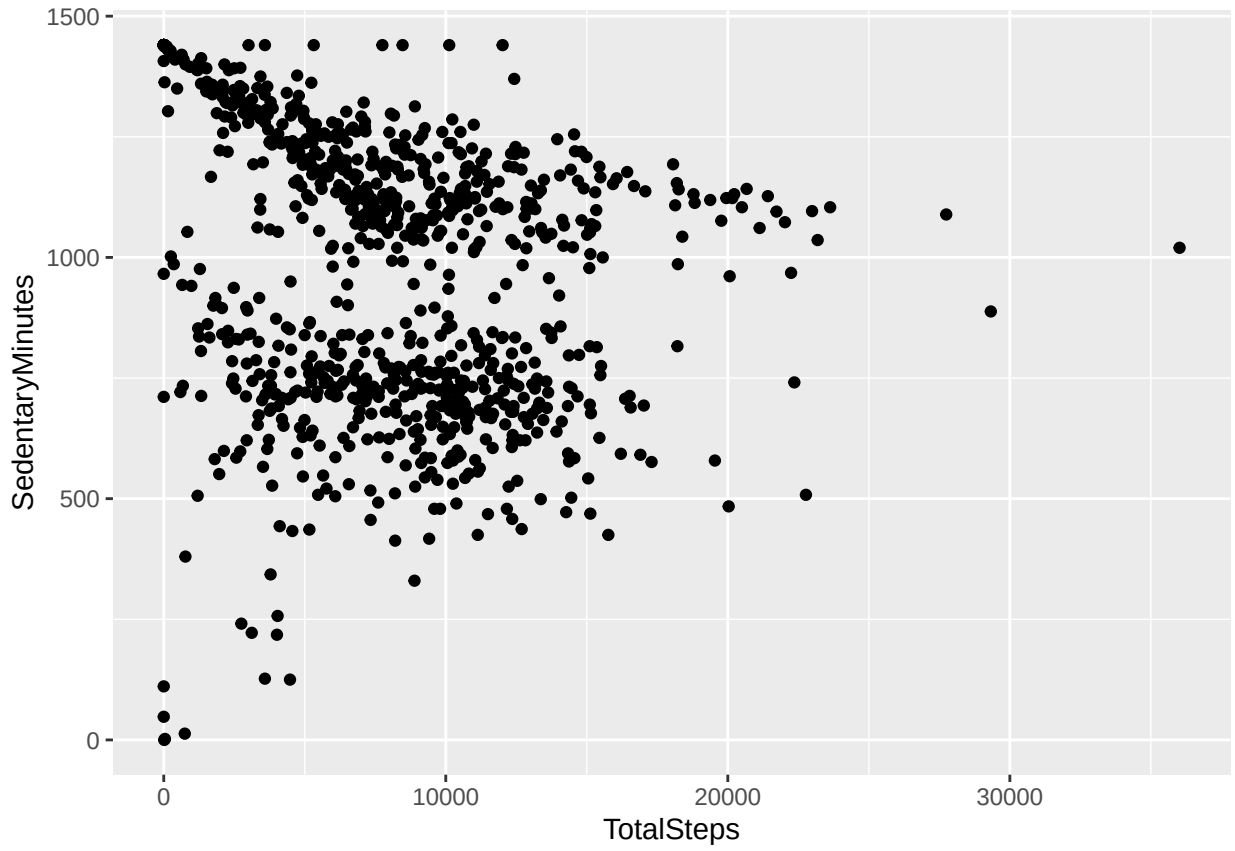
```
##  TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
##  Min.   :1.000     Min.   : 58.0      Min.   : 61.0
##  1st Qu.:1.000     1st Qu.:361.0      1st Qu.:403.0
##  Median :1.000     Median :433.0      Median :463.0
##  Mean   :1.119     Mean   :419.5      Mean   :458.6
##  3rd Qu.:1.000     3rd Qu.:490.0      3rd Qu.:526.0
##  Max.   :3.000     Max.   :796.0      Max.   :961.0
```

What does this tell us about how this sample of people's activities?

Plotting a few explorations

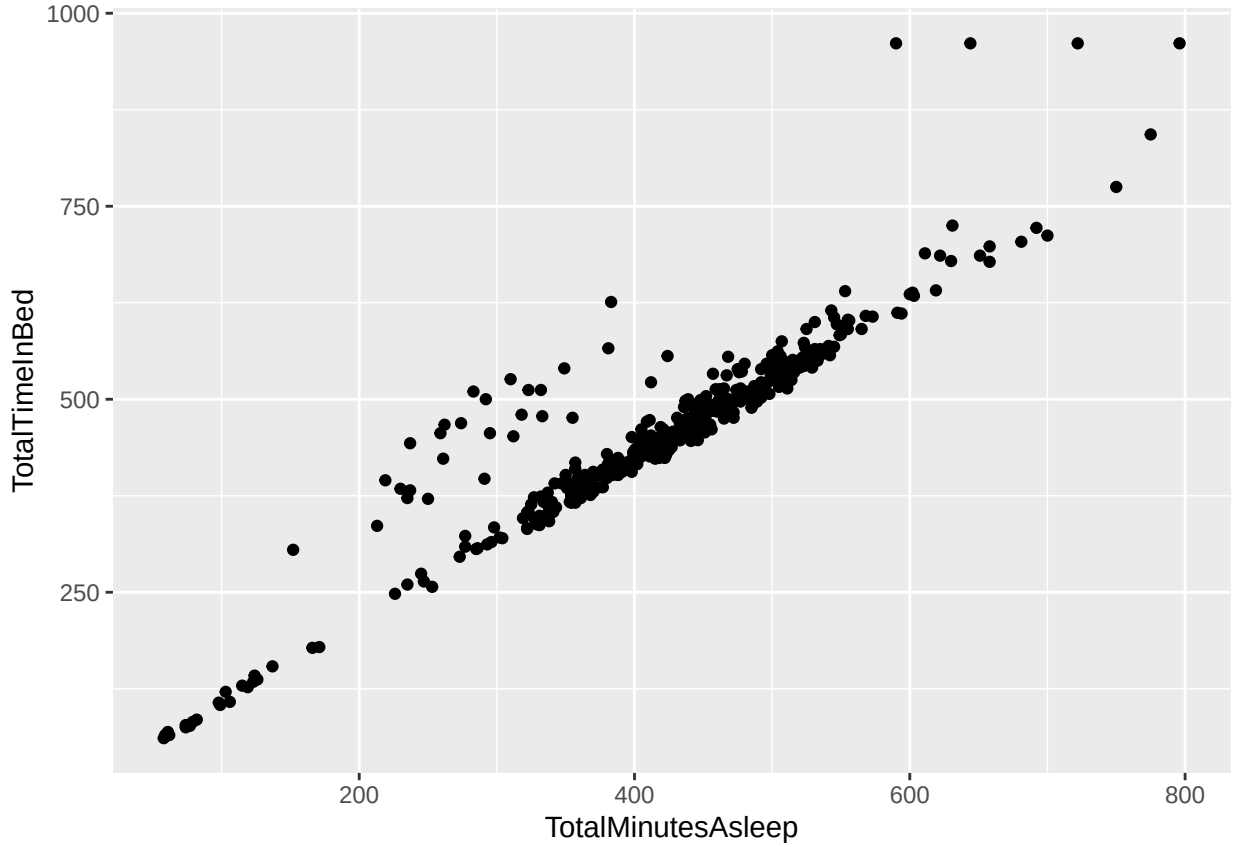
What's the relationship between steps taken in a day and sedentary minutes? How could this help inform the customer segments that we can market to? E.g. position this more as a way to get started in walking more? Or to measure steps that you're already taking?

```
ggplot(data=daily_activity, aes(x=TotalSteps, y=SedentaryMinutes)) + geom_point()
```



What's the relationship between minutes asleep and time in bed? You might expect it to be almost completely linear - are there any unexpected trends?

```
ggplot(data=sleep_day, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) + geom_point()
```



What could these trends tell you about how to help market this product? Or areas where you might want to explore further?

Merging these two datasets together

```
combined_data <- merge(sleep_day, daily_activity, by="Id")
```

Take a look at how many participants are in this data set.

```
n_distinct(combined_data$Id)
```

```
## [1] 24
```

Note that there were more participant Ids in the daily activity dataset that have been filtered out using merge. Consider using 'outer_join' to keep those in the dataset.

Now you can explore some different relationships between activity and sleep as well. For example, do you think participants who sleep more also take more steps or fewer steps per day? Is there a relationship at all? How could these answers help inform the marketing strategy of how you position this new product?

This is just one example of how to get started with this data - there are many other files and questions to explore as well!