```
##############################
## 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 and load the tidyverse
install.packages('tidyverse')
library(tidyverse)

######################
## Load your CSV files ##
######################

# 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")
```

```
# Create another dataframe for the sleep data.
sleep_day <- read.csv("sleepDay_merged.csv")

#############################
## Explore a few key tables ##
#############################

# Take a look at the daily_activity data.

head(daily_activity)


# Identify all the columns in the daily_activity data.
colnames(daily_activity)


# Take a look at the sleep_day data.

head(sleep_day)


# Identify all the columns in the daily_activity data.

colnames(sleep_day)


# 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$Id)
n_distinct(sleep_day$Id)


# How many observations are there in each dataframe?

nrow(daily_activity)
```

```
nrow(sleep_day)


# 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()


# For the sleep dataframe:

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


# 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)


# 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!
```