Process

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

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

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

sleep <- read.csv("sleepDay1_merged.csv") #Sleep data with time removed from
the SleepDay column.
```

Taking a look at the data.

```
head(daily_activity)
```

Id	ActivityDate	TotalSteps	TotalDistance	TrackerDistance
1 1503960366	04-Apr-16	13162	8.50	8.50
2 1503960366	13-Apr-16	10735	6.97	6.97
3 1503960366	14-Apr-16	10460	6.74	6.74
4 1503960366	15-Apr-16	9762	6.28	6.28
5 1503960366	16-Apr-16	12669	8.16	8.16
6 1503960366	17-Apr-16	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	**FAM***
1 6.06	0	25	13
2 4.71	0	21	19
3 3.91	0	30	11
4 2.83	0	29	34
5 5.04	0	36	10
6 2.51	0	38	20

LightlyActiveMinutes	SedentaryMinutes	Calories
1 328	728	1985
2 217	776	1797
3 181	1218	1776
4 209	726	1745
5 221	773	1863
6 164	539	1728

***Fairly Active Minutes**

```
head(sleep_day)
```

Id	SleepDay	TotalSleepRecords	TotalMinutesAsleep	TotalTimeInBed
1 1503960366	12/04/2016 00:00	1	327	346
2 1503960366	13/04/2016 00:00	2	384	407
3 1503960366	15/04/2016 00:00	1	412	442
4 1503960366	16/04/2016 00:00	2	340	367
5 1503960366	17/04/2016 00:00	1	700	712
6 1503960366	19/04/2016 00:00	1	304	32

```
head(sleep)
Id           SleepDay    TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
1 1503960366 12/04/2016  1                 327                346
2 1503960366 13/04/2016  2                 384                407
3 1503960366 15/04/2016  1                 412                442
4 1503960366 16/04/2016  2                 340                367
5 1503960366 17/04/2016  1                 700                712
6 1503960366 19/04/2016  1                 304                320
```

Identify all the columns in the daily_activity data.

```
colnames(daily_activity)

[1] "Id" "ActivityDate" "TotalSteps"
[4] "TotalDistance" "TrackerDistance" "LoggedActivitiesDistance"
[7] "VeryActiveDistance" "ModeratelyActiveDistance" "LightActiveDistance"
[10] "SedentaryActiveDistance" "VeryActiveMinutes" "FairlyActiveMinutes"
[13] "LightlyActiveMinutes" "SedentaryMinutes" "Calories"

colnames(sleep_day)
[1] "Id" "SleepDay" "TotalSleepRecords"
[4] "TotalMinutesAsleep" "TotalTimeInBed"


colnames(sleep)

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

How many unique participants are there in each dataframe?

```
n_distinct(daily_activity$Id) [1] 33
n_distinct(sleep_day$Id) [1] 24
n_distinct(sleep$Id) [1] 24
```

How many observations are there in each dataframe?

```
nrow(daily_activity) [1] 936
nrow(sleep_day) [1] 411
nrow(sleep)  [1] 411
```

Checking columns are correctly formatted.

```
str(sleep)

'data.frame':      411 obs. of  5 variables:
 $ Id                : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
 $ SleepDay          : chr  "12/04/2016" "13/04/2016" "15/04/2016"
"16/04/2016" ...
 $ TotalSleepRecords : int  1 2 1 2 1 1 1 1 1 1 ...
 $ TotalMinutesAsleep: int  327 384 412 340 700 304 360 325 361 430 ...
 $ TotalTimeInBed    : int  346 407 442 367 712 320 377 364 384 449 ...
```

```
str(sleep_day)
```

```
'data.frame':	411 obs. of  5 variables:
 $ Id                : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
 $ SleepDay          : chr  "12/04/2016 00:00" "13/04/2016 00:00" "15/04/2016
00:00" "16/04/2016 00:00" ...
 $ TotalSleepRecords : int  1 2 1 2 1 1 1 1 1 1 ...
 $ TotalMinutesAsleep: int  327 384 412 340 700 304 360 325 361 430 ...
 $ TotalTimeInBed    : int  346 407 442 367 712 320 377 364 384 449 ...
```

```
str(daily_activity)
```

```
'data.frame':	936 obs. of  13 variables:
 $ Id                       : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
 $ Date                     : Date, format: "2016-04-04" "2016-04-13" ...
 $ TotalSteps               : int  13162 10735 10460 9762 12669 9705 13019
15506 10544 9819 ...
 $ TotalDistance            : num  8.5 6.97 6.74 6.28 8.16 ...
 $ VeryActiveDistance       : num  1.88 1.57 2.44 2.14 2.71 ...
 $ ModeratelyActiveDistance: num  0.55 0.69 0.4 1.26 0.41 ...
 $ LightActiveDistance      : num  6.06 4.71 3.91 2.83 5.04 ...
 $ SedentaryActiveDistance : num  0 0 0 0 0 0 0 0 0 0 ...
 $ VeryActiveMinutes        : int  25 21 30 29 36 38 42 50 28 19 ...
 $ FairlyActiveMinutes      : int  13 19 11 34 10 20 16 31 12 8 ...
 $ LightlyActiveMinutes     : int  328 217 181 209 221 164 233 264 205 211 ...
 $ SedentaryMinutes         : int  728 776 1218 726 773 539 1149 775 818 838
...
 $ Calories                 : int  1985 1797 1776 1745 1863 1728 1921 2035 1786
1775 ...
```

Quick summary statistics we want to know about each data frame.

For the sleep day data frame:

```
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.:402.5
Median :1.000     Median :432.0      Median :463.0
Mean :1.119       Mean :419.1        Mean :458.3
3rd Qu.:1.000     3rd Qu.:490.0      3rd Qu.:526.0
Max. :3.000       Max. :796.0        Max. :961.0
```

For the daily activity dataframe:

```
daily_activity %>%
select(TotalSteps,TotalDistance,TrackerDistance,LoggedActivitiesDistance, +
VeryActiveDistance,ModeratelyActiveDistance,LightActiveDistance, +
SedentaryActiveDistance,VeryActiveMinutes,FairlyActiveMinutes, +
LightlyActiveMinutes,SedentaryMinutes,Calories) %>%
summary()
```

```
TotalSteps      TotalDistance   TrackerDistance  LoggedActivitiesDistance
Min. : 0        Min. : 0.000    Min. : 0.000     Min. :0.0000
1st Qu.: 3818   1st Qu.: 2.645  1st Qu.: 2.645   1st Qu.:0.0000
Median : 7441   Median : 5.265  Median : 5.265   Median :0.0000
Mean : 7671     Mean : 5.513    Mean : 5.499     Mean :0.1086
3rd Qu.:10734   3rd Qu.: 7.720  3rd Qu.: 7.713   3rd Qu.:0.0000
Max. :36019     Max. :28.030    Max. :28.030     Max. :4.9421

VeryActiveDistance   ModeratelyActiveDistance   LightActiveDistance
Min. : 0.000         Min. :0.00                 Min. : 0.000
1st Qu.: 0.000       1st Qu.:0.00               1st Qu.: 1.960
Median : 0.220       Median :0.24               Median : 3.380
Mean : 1.509         Mean :0.57                 Mean : 3.355
3rd Qu.: 2.090       3rd Qu.:0.80               3rd Qu.: 4.790
Max. :21.920         Max. :6.48                 Max. :10.710

SedentaryActiveDistance VeryActiveMinutes FairlyActiveMinutes LAM*
Min. :0.000000          Min. : 0.00       Min. : 0.00         Min. : 0.0
1st Qu.:0.000000        1st Qu.: 0.00     1st Qu.: 0.00       1st Qu.:128.0
Median :0.000000        Median : 4.00     Median : 7.00       Median :199.0
Mean :0.001613          Mean : 21.26      Mean : 13.62        Mean :193.6
3rd Qu.:0.000000        3rd Qu.: 32.00    3rd Qu.: 19.00      3rd Qu.:264.2
Max. :0.110000          Max. :210.00      Max. :143.00        Max. :518.0

SedentaryMinutes  Calories
Min. : 0.0        Min. : 52
1st Qu.: 729.0    1st Qu.:1834
Median :1057.0    Median :2144
Mean : 989.3      Mean :2313
3rd Qu.:1226.0    3rd Qu.:2794
Max. :1440.0      Max. :4900
```

***Lightly Active Minutes**

Changing the format of ActivityDate and SleepDay to 'Date'

```
daily_activity$ActivityDate <- as.Date(daily_activity$ActivityDate, format=
"%d/%m/%Y" )

sleep$SleepDay <- as.Date(sleep$SleepDay, format= "%d/%m/%Y")
```

Removing unnecessary columns from the data frame

```
daily_activity <- daily_activity %>%
  select(-"TrackerDistance", -"LoggedActivitiesDistance")
```

Matching the date columns for conformance consistency

```
daily_activity <- rename(daily_activity, Date = ActivityDate)
sleep <- rename(sleep, Date = SleepDay)
```

Merging the two datasets by 'Id' and 'Date'

```
combined_data <- inner_join(daily_activity, sleep, by = c("Id", "Date"))

n_distinct(combined_data$Id)

[1] 24
```

Adding a 'Time Awake' column - to see how much time users are spending when they're not asleep

```
combined_data <- combined_data %>%
  mutate(TimeAwake = TotalTimeInBed - TotalMinutesAsleep)
```

Adding 'Total Active minutes' column

```
combined_data <- combined_data %>%
  mutate(TotalActiveMinutes = VeryActiveMinutes + FairlyActiveMinutes +
LightlyActiveMinutes)
```

Adding a 'day of week' column

```
combined_data$day_of_week <- format(as.Date(combined_data$Date), "%a")

combined_data$day_of_week <- ordered(combined_data$day_of_week,levels=c("Mon",
"Tue", "Wed", "Thu", "Fri", "Sat","Sun"))
```

Removing duplicates

```
combined_data <- distinct(combined_data)
```

Analysis

```
summary(combined_data)

       Id                 Date              TotalSteps    TotalDistance
 Min.   :1.504e+09   Min.   :2016-04-12   Min.   :   17   Min.   : 0.010
 1st Qu.:3.977e+09   1st Qu.:2016-04-19   1st Qu.: 5178   1st Qu.: 3.587
 Median :4.703e+09   Median :2016-04-27   Median : 8913   Median : 6.270
 Mean   :4.995e+09   Mean   :2016-04-26   Mean   : 8506   Mean   : 6.006
 3rd Qu.:6.962e+09   3rd Qu.:2016-05-04   3rd Qu.:11335   3rd Qu.: 7.975
 Max.   :8.792e+09   Max.   :2016-05-12   Max.   :22770   Max.   :17.540

VeryActiveDistance ModeratelyActiveDistance LightActiveDistance
 Min.   : 0.000     Min.   :0.0000           Min.   :0.010
 1st Qu.: 0.000     1st Qu.:0.0000           1st Qu.:2.540
 Median : 0.565     Median :0.4150           Median :3.655
 Mean   : 1.447     Mean   :0.7445           Mean   :3.784
 3rd Qu.: 2.380     3rd Qu.:1.0400           3rd Qu.:4.912
 Max.   :12.540     Max.   :6.4800           Max.   :9.480
```

```
SedentaryActiveDistance VeryActiveMinutes FairlyActiveMinutes  LAM*
 Min.   :0.0000000       Min.   :  0.00    Min.   :  0.00      Min.   :  2.0
 1st Qu.:0.0000000       1st Qu.:  0.00    1st Qu.:  0.00      1st Qu.:158.0
 Median :0.0000000       Median :  8.50    Median : 11.00      Median :208.0
 Mean   :0.0009314       Mean   : 24.95    Mean   : 17.94      Mean   :216.4
 3rd Qu.:0.0000000       3rd Qu.: 38.00    3rd Qu.: 27.00      3rd Qu.:263.0
 Max.   :0.1100000       Max.   :210.00    Max.   :143.00      Max.   :518.0

 SedentaryMinutes    Calories    TotalSleepRecords TotalMinutesAsleep
 Min.   :   0.0   Min.   : 257   Min.   :1.00      Min.   : 58.0
 1st Qu.: 630.2   1st Qu.:1838   1st Qu.:1.00      1st Qu.:362.5
 Median : 717.0   Median :2207   Median :1.00      Median :433.0
 Mean   : 712.0   Mean   :2387   Mean   :1.12      Mean   :419.6
 3rd Qu.: 783.2   3rd Qu.:2912   3rd Qu.:1.00      3rd Qu.:490.5
 Max.   :1265.0   Max.   :4900   Max.   :3.00      Max.   :796.0

 TotalTimeInBed    TimeAwake      TotalActiveMinutes day_of_week
 Min.   : 61.0   Min.   :  0.00   Min.   :  2.0      Mon:46
 1st Qu.:406.0   1st Qu.: 17.00   1st Qu.:206.0      Tue:63
 Median :463.5   Median : 26.00   Median :263.5      Wed:66
 Mean   :459.0   Mean   : 39.41   Mean   :259.3      Thu:64
 3rd Qu.:526.2   3rd Qu.: 40.00   3rd Qu.:314.5      Fri:57
 Max.   :961.0   Max.   :371.00   Max.   :540.0      Sat:57
                                                     Sun:55
```

***Lightly Active Minutes**

Average number of steps by they of week

```
combined_data %>%
  select(TotalSteps, day_of_week)%>%
  group_by(day_of_week)%>%
  summarise(TotalSteps= round(c(TotalSteps= mean(TotalSteps)),0))
```

```
# A tibble: 7 × 2
  day_of_week    TotalSteps
  <ord>            <dbl>
1 Mon               9273
2 Tue               9144
3 Wed               8023
4 Thu               8184
5 Fri               7901
6 Sat               9871
7 Sun               7298
```

Average distance by day of week

```
combined_data %>%
  select(TotalDistance, day_of_week)%>%
  group_by(day_of_week)%>%
  summarise(TotalDistance= round(c(TotalDistance= mean(TotalDistance)),2))
```

```
# A tibble: 7 × 2
  day_of_week    TotalDistance
  <ord>               <dbl>
1 Mon                  6.54
2 Tue                  6.4
3 Wed                  5.72
4 Thu                  5.77
5 Fri                  5.51
6 Sat                  7.02
7 Sun                  5.18
```

Average hours of sleep by day of week

```
combined_data %>%
  select(TotalMinutesAsleep, day_of_week)%>%
  group_by(day_of_week)%>%
  summarise(TotalMinutesAsleep= round(c(mean(TotalMinutesAsleep)/60),2))
```

```
# A tibble: 7 × 2
  day_of_week      TotalMinutesAsleep
  <ord>                         <dbl>
1 Mon                            6.99
2 Tue                            6.78
3 Wed                            7.24
4 Thu                            6.69
5 Fri                            6.76
6 Sat                            6.98
7 Sun                            7.55
```

Average minutes of activity by intensity

```
combined_data %>%
  summarise(intensity= ordered(c("Very Active", "Fairly Active", "Lightly
Active", "Sedentary Time"), levels=c("Very Active", "Fairly Active", "Lightly
Active", "Sedentary Time")), avg_minutes = round(c(VeryActiveMinutes=
mean(VeryActiveMinutes), FairlyActiveMinutes= mean(FairlyActiveMinutes),
LightlyActiveMinutes= mean(LightlyActiveMinutes),
SedentaryMinutes= mean(SedentaryMinutes)),0))
```

```
     intensity       avg_minutes
1    Very Active          25
2  Fairly Active          18
3 Lightly Active         216
4 Sedentary Time         712
```

Average calories burnt by day of week

```
combined_data %>%
  select(Calories, day_of_week)%>%
  group_by(day_of_week)%>%
  summarise(Calories= round(c(mean(Calories)),0))
```

```
# A tibble: 7 × 2
  day_of_week   Calories
  <ord>          <dbl>
1 Mon             2432
2 Tue             2486
3 Wed             2378
4 Thu             2307
5 Fri             2330
6 Sat             2507
7 Sun             2277
```

How many users are walking more than 7k steps

```
combined_data %>%
  select(Id, TotalSteps) %>%
  group_by(Id) %>%
  summarise(meansteps = mean(TotalSteps))%>%
  filter(meansteps >= 7000)
```

```
# A tibble: 15 × 2
          Id     meansteps
        <dbl>     <dbl>
 1 1503960366    12374.
 2 1644430081     7968.
 3 2347167796     8533.
 4 3977333714    11218
 5 4319703577     7125.
 6 4388161847    11034.
 7 4558609924     8139
 8 4702921684     9036.
 9 5553957443     8613.
10 5577150313     9260.
11 6117666160     8824.
12 6962181067     9795.
13 7086361926    10290.
14 8053475328    19079.
15 8378563200     8754.
```

Share - Visualisations:

```
ggplot(data=sleep_day, aes(
  x=TotalMinutesAsleep,
  y=TotalTimeInBed)) + geom_point() +
  geom_smooth(method = 'loess')+
  labs(title = "Time in Bed vs Minutes Asleep",
       x = 'Total Minutes Asleep', y = 'Total Time in Bed', color =
'TimeInBed') +
 theme(plot.title = element_text(size = 14,face = "bold"))
```

```
ggplot(data= daily_activity, aes(
 x = TotalSteps,
 y = Calories,
color = Calories)) + geom_point() +
scale_color_gradient(low="black", high="blue") + geom_smooth(method =
'loess')+  labs(title= 'Calories vs Steps',
 x = 'Total Steps', y = 'Daily Calories', color = 'Calories')+
theme(plot.title = element_text(size = 14, face = "bold"))
```

```
ggplot(data= daily_activity, aes(
  x = VeryActiveMinutes,
  y = Calories,
  color = Calories)) + geom_point() +
  scale_color_gradient(low="red", high="blue4") + geom_smooth(method =
'loess')+
  labs(title= 'Intense Activity vs Calories',
       x = 'Intense Actvity', y = 'Calories', color = 'Calories')+
       theme(plot.title = element_text(size = 14, face = "bold")
```

```
ggplot(data =daily_activity, aes(
  x = FairlyActiveMinutes,
  y = Calories,
  color = Calories)) + geom_point() +
  scale_color_gradient(low="red", high="blue4") + geom_smooth(method =
'loess')+
  labs(title= 'Moderate Activity vs Calories',
       x = 'Moderate Actvity', y = 'Calories', color = 'Calories')+
  theme(plot.title = element_text(size = 14, face = "bold"))
```

```
ggplot(data = daily_activity, aes(
       X = LightlyActiveMinutes,
       Y = Calories,
       color = Calories)) + geom_point() +
scale_color_gradient(low= “red”, high= “blue4”) + geom_smooth(method =
‘loess’)+
labs(title= ‘Light Activity vs Calories’,
     X = ‘Light Activity’, Y = ‘Calories’, color = ‘Calories’) +
theme(plot.title = element_text(size = 14, face = “bold”))
```

```
ggplot(data = daily_activity,aes(
       X = SedentaryMinutes,
       Y = Calories,
       colour = Calories)) + geom_point() +
scale_colour_gradient(low= “red”, high= “blue4”) + geom_smooth(method =
```

```
‘loess’) +
Labs(title = ‘Sedentary Activity vs Calories’,
     X = ‘Sedentary Activity’, y = ‘Calories’, color = ‘Calories’) +
Theme(plot.title = element_text(size = 14, face = “bold”))
```

```
ggplot(data=daily_activity,aes
       (x= TotalDistance,
        y= Calories, color = Calories)) + geom_point()+
  scale_color_gradient(low= "deeppink", high= "darkred") + geom_smooth(method
= 'loess') +
  labs(title="Calories burned by distance",
       x = "Total Distance",y = "Calories",color = 'Calories') +
 theme(plot.title = element_text(size = 14,face = "bold"))
```

```
combined_data %>%
  select(TotalSteps, day_of_week)%>%
  group_by(day_of_week)%>%
  summarise(TotalSteps= round(c(TotalSteps= mean(TotalSteps)),0))%>%
  ggplot(aes(x=day_of_week, y=TotalSteps, fill=TotalSteps)) +
geom_bar(stat='identity') +
  labs(title = "Average Steps by Weekday", subtitle="",
       x="Weekday", y="Average steps", fill="Average Steps") +
  geom_hline(yintercept=7000, linetype="dashed", color = "red")+
  geom_text(aes(x=day_of_week, y= TotalSteps, label = TotalSteps),
            vjust = -0.5, size = 2)+
  annotate("text", x='Thu', y=7400, label= "7k steps", size= 2, color="red4")
```

```
combined_data %>%
  select(TotalDistance, day_of_week)%>%
  group_by(day_of_week)%>%
  summarise(TotalDistance= round(c(TotalDistance= mean(TotalDistance)),2))%>%
  ggplot(aes(x=day_of_week, y=TotalDistance, fill=TotalDistance)) +
geom_bar(stat='identity')+
  labs(title = "Average distance(km) by Weekday", subtitle="",
       x="Weekday", y="Average distance(km)", fill="Average distance") +
  geom_text(aes(x=day_of_week, y= TotalDistance, label = TotalDistance),
            vjust = -0.5, size = 2)
```

```
combined_data %>%
select(TotalMinutesAsleep, day_of_week)%>%
group_by(day_of_week)%>%
summarise(TotalMinutesAsleep= round(c(mean(TotalMinutesAsleep)/60),2)) %>%
ggplot(aes(x=day_of_week, y=TotalMinutesAsleep, fill=TotalMinutesAsleep)) +
geom_bar(stat='identity')+
labs(title = "Average hours of sleep by Weekday", subtitle="",
x="", y="Average sleep", fill="Average Sleep") +
geom_hline(yintercept=7, linetype="dashed", color = "red4") +
geom_text(aes(x=day_of_week, y= TotalMinutesAsleep, label =
TotalMinutesAsleep),
vjust = -0.5, size = 2) +
annotate("text", x='Sun', y=7.2, label= "7 hours", size= 2, color="red4")
```