Documentacion Etapa IV - Analisis

Emiliano Garcia

2025-08-12

Etapa IV - Analisis

```
library(tidyverse)
library(corrplot)
master_table_clean <- read_csv ("master_table_clean.csv")
```

Previsualizacion de los datos

```
head (master_table_clean)
```

```
## # A tibble: 6 x 8
##           Id ActivityDate   TotalSteps TotalDistance Calories TotalMinutesAsleep
##        <dbl> <chr>               <dbl>         <dbl>    <dbl>               <dbl>
## 1 1503960366 2016-03-25 00~      11004          7.11     1819                  NA
## 2 1503960366 2016-03-26 00~      17609         11.6      2154                  NA
## 3 1503960366 2016-03-27 00~      12736          8.53     1944                  NA
## 4 1503960366 2016-03-28 00~      13231          8.93     1932                  NA
## 5 1503960366 2016-03-29 00~      12041          7.85     1886                  NA
## 6 1503960366 2016-03-30 00~      10970          7.16     1820                  NA
## # i 2 more variables: WeightKg <dbl>, SourcePeriod <chr>
```

```
colnames (master_table_clean)
```

```
## [1] "Id"                 "ActivityDate"       "TotalSteps"
## [4] "TotalDistance"      "Calories"           "TotalMinutesAsleep"
## [7] "WeightKg"           "SourcePeriod"
```

```
str (master_table_clean)
```

```
## spc_tbl_ [1,373 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id                : num [1:1373] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDate      : chr [1:1373] "2016-03-25 00:00:00.000000 UTC" "2016-03-26 00:00:00.000000 UTC
##  $ TotalSteps        : num [1:1373] 11004 17609 12736 13231 12041 ...
##  $ TotalDistance     : num [1:1373] 7.11 11.55 8.53 8.93 7.85 ...
##  $ Calories          : num [1:1373] 1819 2154 1944 1932 1886 ...
##  $ TotalMinutesAsleep: num [1:1373] NA NA NA NA NA NA NA NA NA NA ...
```

```
##  $ WeightKg          : num [1:1373] NA NA NA NA NA NA NA NA NA NA ...
##  $ SourcePeriod      : chr [1:1373] "03-04" "03-04" "03-04" "03-04" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   ActivityDate = col_character(),
##   ..   TotalSteps = col_double(),
##   ..   TotalDistance = col_double(),
##   ..   Calories = col_double(),
##   ..   TotalMinutesAsleep = col_double(),
##   ..   WeightKg = col_double(),
##   ..   SourcePeriod = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
summary (master_table_clean)
```

```
##        Id            ActivityDate         TotalSteps    TotalDistance
##  Min.   :1.504e+09   Length:1373        Min.   :    0   Min.   : 0.000
##  1st Qu.:2.320e+09   Class :character   1st Qu.: 3325   1st Qu.: 2.290
##  Median :4.445e+09   Mode  :character   Median : 7142   Median : 5.030
##  Mean   :4.782e+09                      Mean   : 7465   Mean   : 5.356
##  3rd Qu.:6.962e+09                      3rd Qu.:10688   3rd Qu.: 7.630
##  Max.   :8.878e+09                      Max.   :51072   Max.   :44.360
##
##     Calories    TotalMinutesAsleep    WeightKg      SourcePeriod
##  Min.   :   0   Min.   :  58.0     Min.   : 52.60   Length:1373
##  1st Qu.:1820   1st Qu.: 364.5     1st Qu.: 61.50   Class :character
##  Median :2138   Median : 436.5     Median : 62.50   Mode  :character
##  Mean   :2319   Mean   : 437.3     Mean   : 72.47
##  3rd Qu.:2786   3rd Qu.: 499.0     3rd Qu.: 85.25
##  Max.   :9718   Max.   :1500.0     Max.   :133.50
##                 NA's   :963        NA's   :1275
```

Estadisticas descriptivas – Media, Mediana y Desvio estandar

Tabla de medias por usuario

```
mean_users <- master_table_clean %>%
  group_by (Id) %>%
  summarise (
    mean_steps = mean (TotalSteps),
    mean_distance = mean (TotalDistance),
    mean_calories = mean (Calories)
  )
View (mean_users)
```

Tabla de medianas por usuario

```r
median_users <- master_table_clean %>%
  group_by (Id) %>%
  summarise (
    median_steps = median (TotalSteps),
    median_distance = median (TotalDistance),
    median_calories = median (Calories)
  )
View (median_users)
```

Tabla de desvios por usuario

```r
sd_users <- master_table_clean %>%
  group_by (Id) %>%
  summarise (
    sd_steps = sd (TotalSteps),
    sd_distance = sd (TotalDistance),
    sd_calories = sd (Calories)
  )
View (sd_users)
```

Graficos para entender distribuciones de datos.

Todas son sesgadas a la derecha, es decir, concentran la mayor cantidad de datos a la izquierda

Grafico I. Distribucion de pasos diarios

```r
ggplot (data = master_table_clean) +
  geom_histogram (mapping = aes (x = TotalSteps), fill = "green") +
  labs (title = "Distribucion de pasos diarios")
```

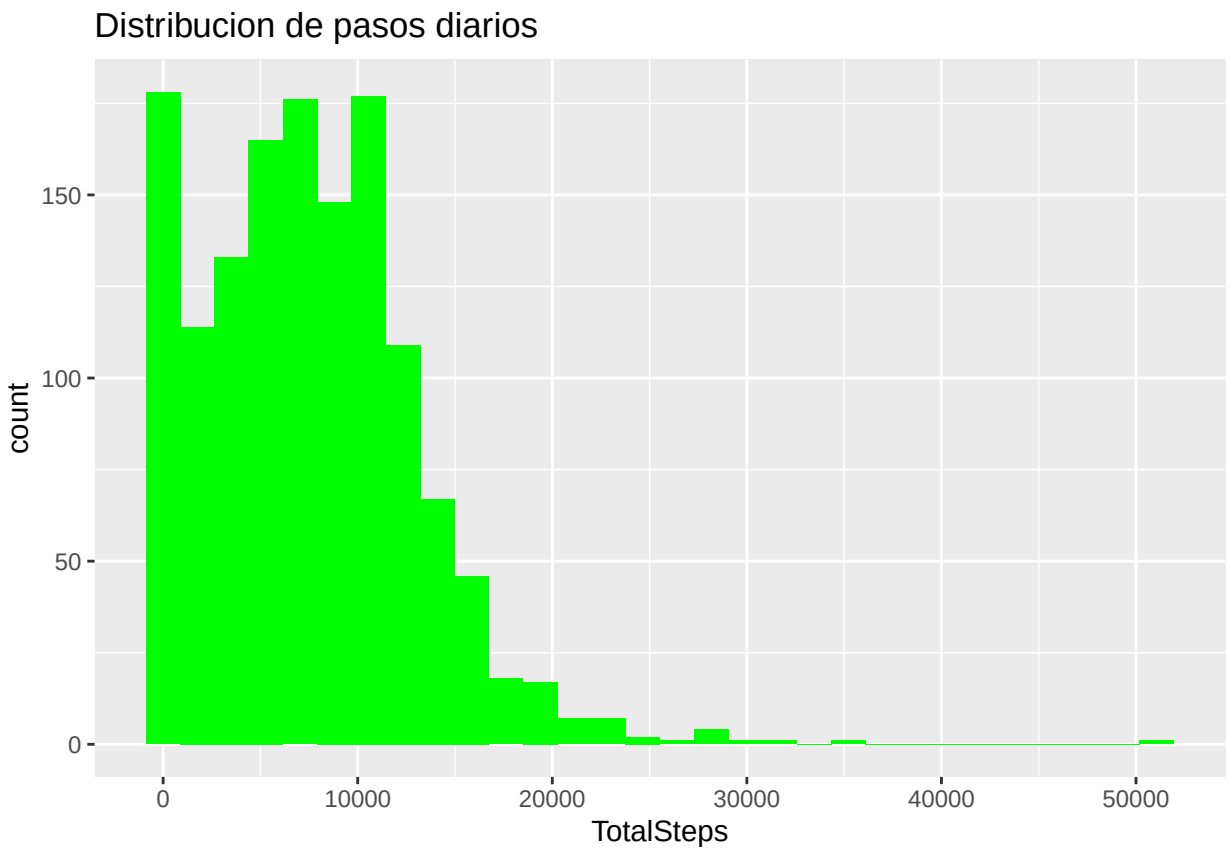


Grafico II. Distribucion de calorias diarias

```
ggplot (data = master_table_clean) +
  geom_histogram (mapping = aes (x = Calories), fill = "orange") +
  labs (title = "Distribucion de calorias diarias")
```

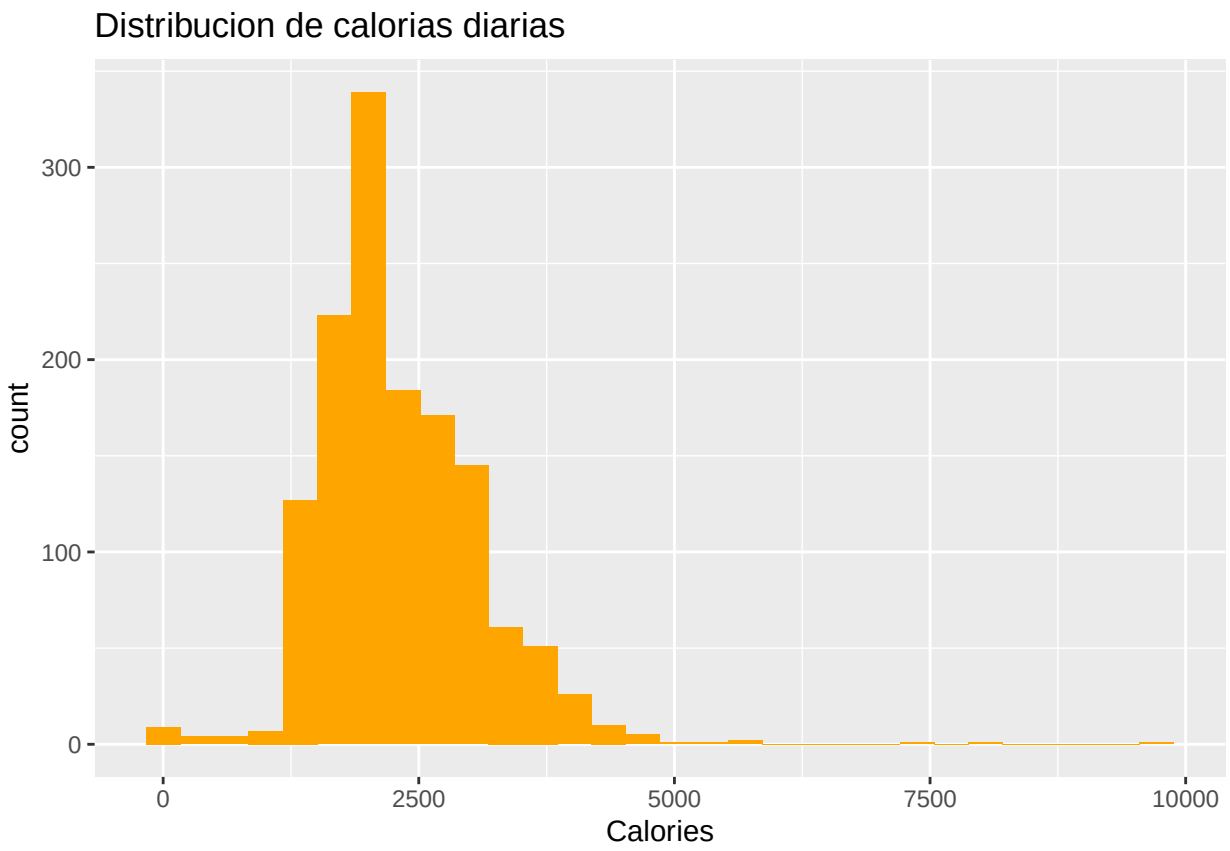
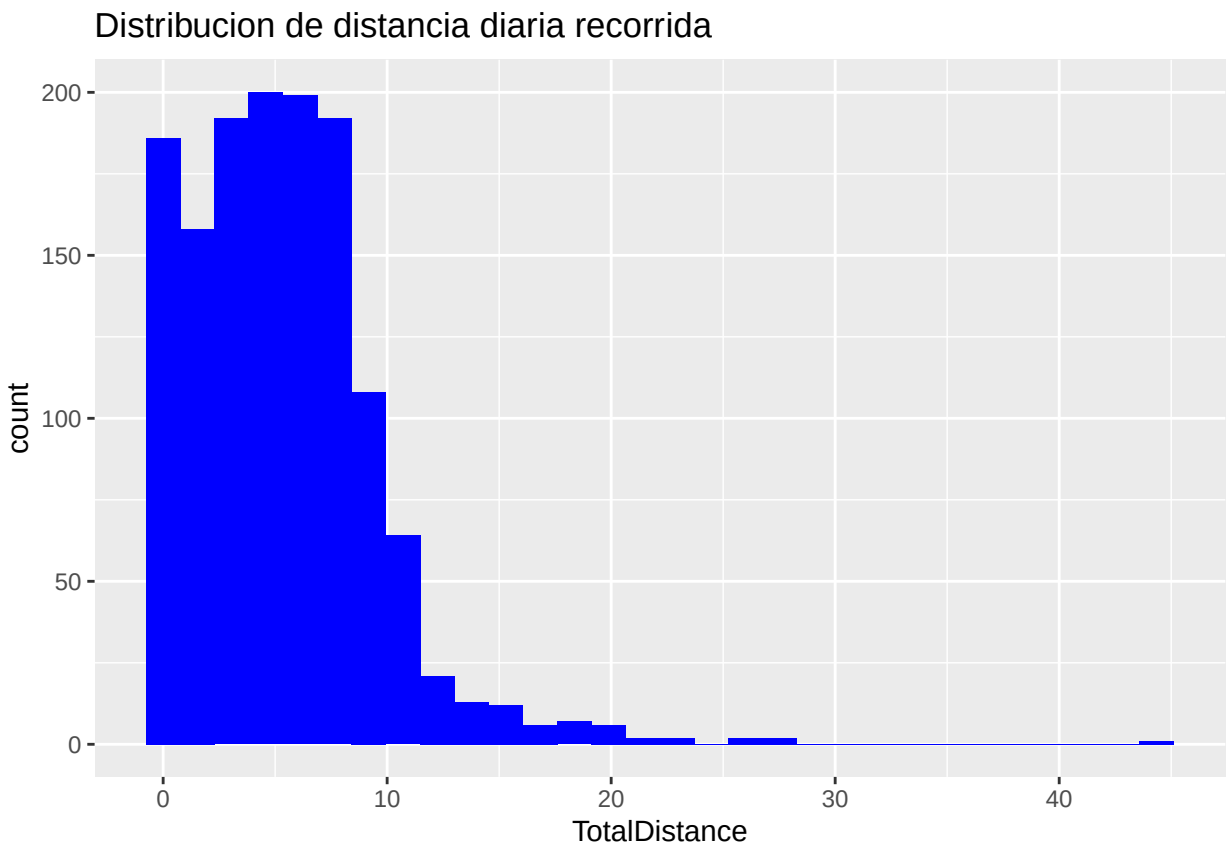


Grafico III. Distribucion de la distancia total recorrida

```
ggplot (data = master_table_clean) +
  geom_histogram (mapping = aes (x = TotalDistance), fill = "blue") +
  labs (title = "Distribucion de distancia diaria recorrida")
```



Graficos para relacionar variables

Matriz de correlaciones

```
var_table <- master_table_clean %>%
 dplyr::select ("TotalSteps", "TotalDistance", "Calories")

cor_matriz <- cor (var_table, use = "complete.obs")
corrplot (cor_matriz, method = "number", type = "upper")
```

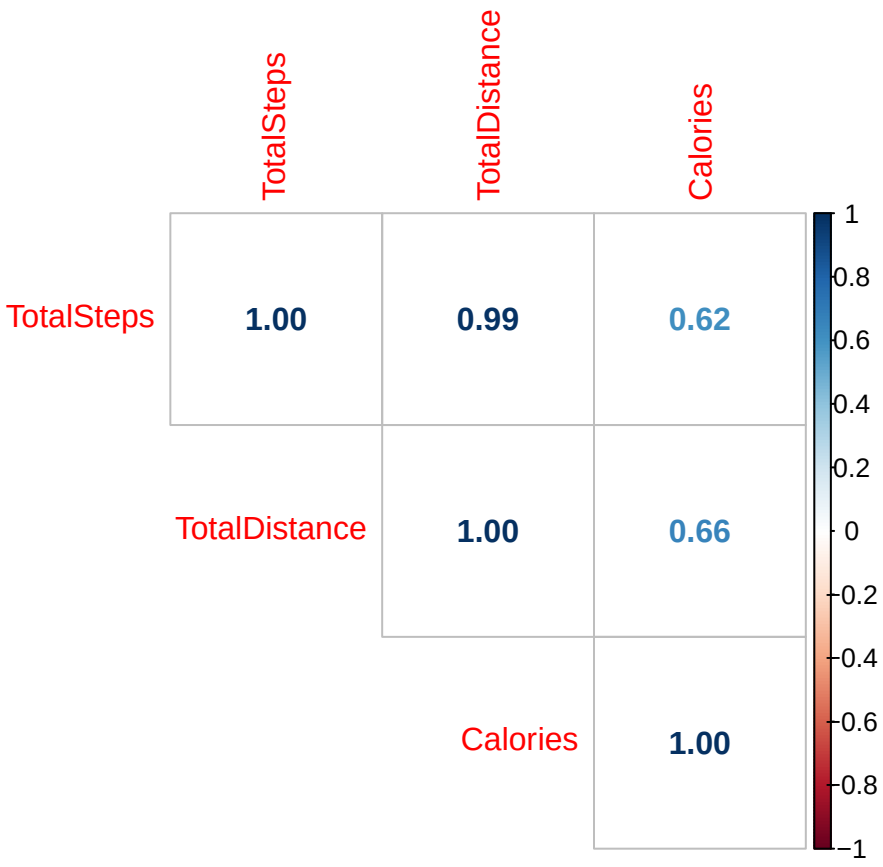
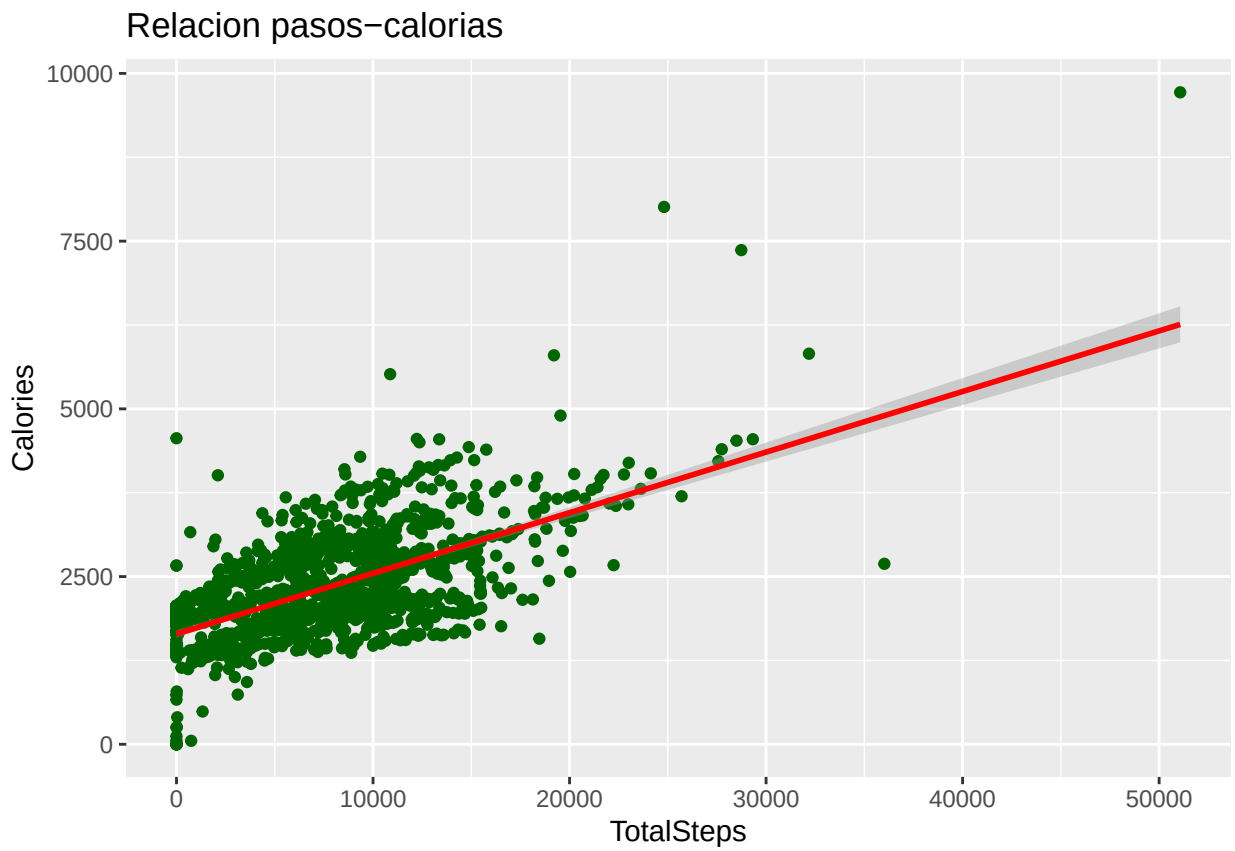


Grafico IV. Correlaciones

```
ggplot (data = master_table_clean, aes (x = TotalSteps, y = Calories)) +
  geom_point (color = "darkgreen") +
  geom_smooth (method = "lm", color = "red") +
  labs (title = "Relacion pasos-calorias")
```



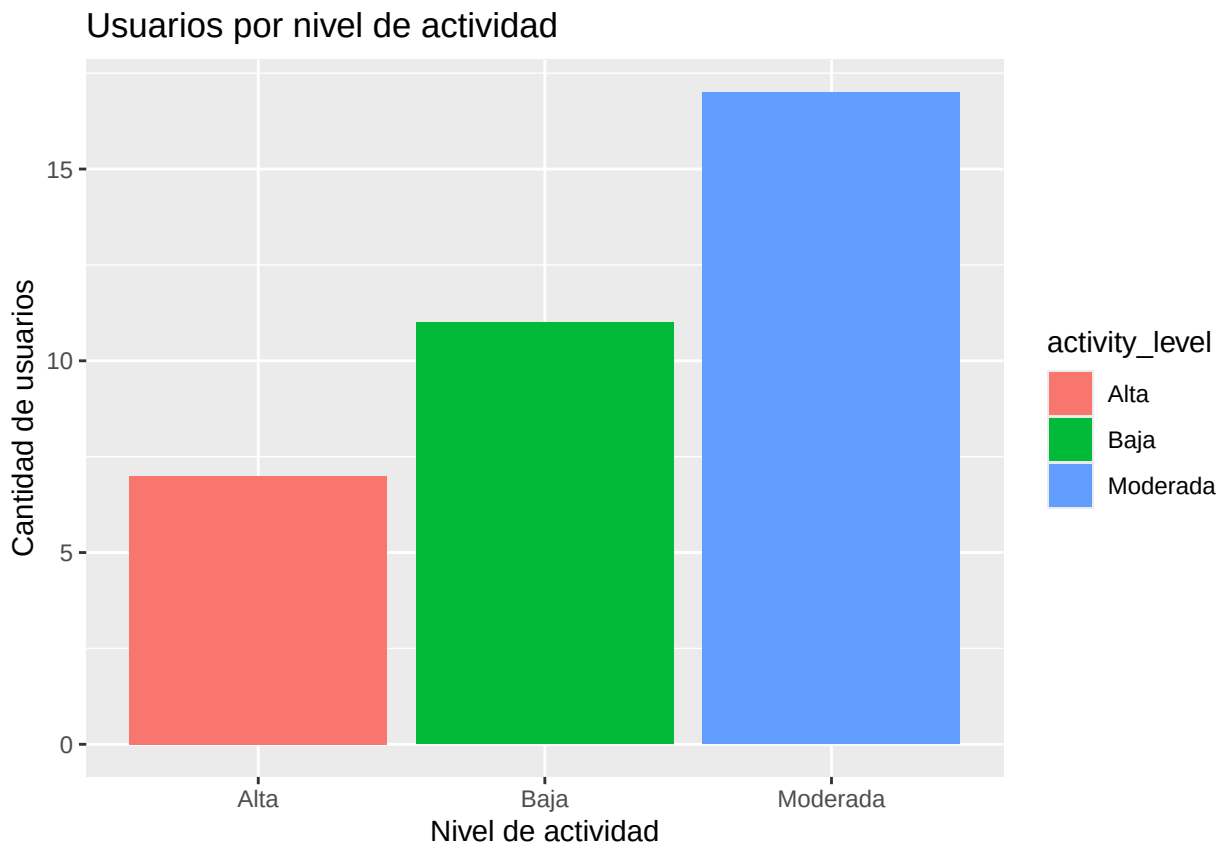
Agrupamos usuarios por nivel de actividad

Tabla segun nivel de actividad

```
activity_table <- master_table_clean %>%
  group_by (Id) %>%
  summarise (mean_steps = mean (TotalSteps)) %>%
  mutate (activity_level = case_when (
    mean_steps < 5000 ~ "Baja",
    mean_steps < 10000 ~ "Moderada",
    TRUE ~ "Alta"
  ))
```

Grafico V. Usuarios por nivel de actividad

```
ggplot (data = activity_table, aes (x = activity_level, fill = activity_level)) +
  geom_bar () +
  labs (title = "Usuarios por nivel de actividad",
        x = "Nivel de actividad",
        y = "Cantidad de usuarios")
```



Actividad segun dia de la semana

Convertimos ActivityDate a formato fecha

```
master_table_clean <- master_table_clean %>%
  mutate (ActivityDate = as.Date(ActivityDate))
```

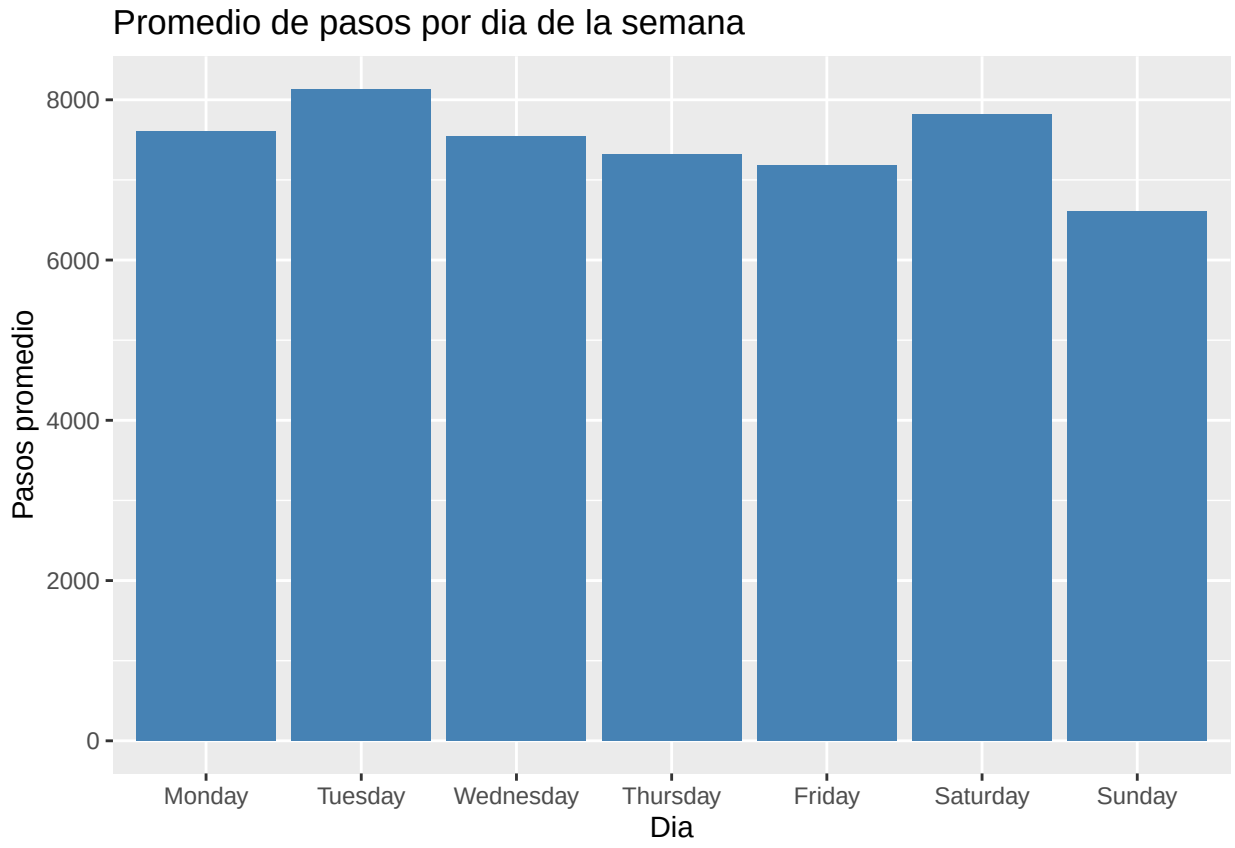
Tabla segun dias

```
weekday_table <- master_table_clean %>%
  mutate (weekday = weekdays(ActivityDate)) %>%
  group_by (weekday) %>%
  summarise (avg_steps = mean (TotalSteps))

weekday_table <- weekday_table %>%
  mutate (weekday = factor (weekday,
                            levels = c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturda
  arrange (weekday)
```

Grafico VI. Promedio de pasos por dia de la semana

```
ggplot (data = weekday_table, aes (x = weekday, y = avg_steps)) +
          geom_col (fill = "steelblue") +
          labs (title = "Promedio de pasos por dia de la semana",
                x = "Dia",
                y = "Pasos promedio")
```



Promedio diario de pasos por usuario, en general, y su evolucion temporal

Tabla

```
daily_trend <- master_table_clean %>%
  group_by (ActivityDate) %>%
  summarise (avg_steps = mean (TotalSteps))
```

El dia con menos pasos totales realizados por los 35 usuarios en su conjunto fue el 2016-03-18, con tan solo 658 pasos registrados por cada uno de ellos, en promedio. Por su parte, El dia con mas pasos totales realizados por el conjunto de usuarios que compartieron sus datos fue el 2016-04-12 con 10.780 para cada uno de ellos, en promedio.

Grafico VII. Tendencia temporal del promedio diario de pasos

```
ggplot (data = daily_trend, aes (x = ActivityDate, y = avg_steps)) +
  geom_line (color = "darkblue") +
  labs (title = "Tendencia del promedio diario de pasos",
        x = "Fecha",
        y = "Pasos promedio")
```

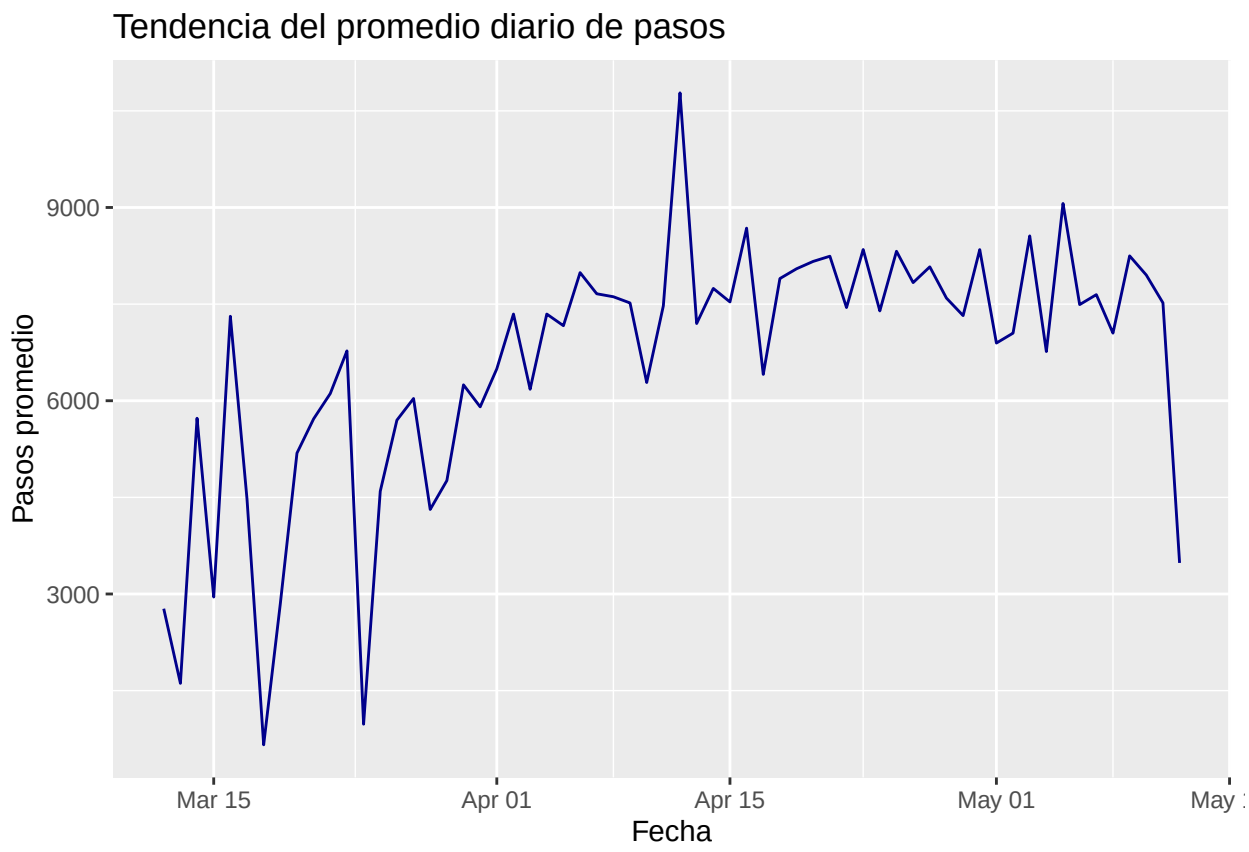
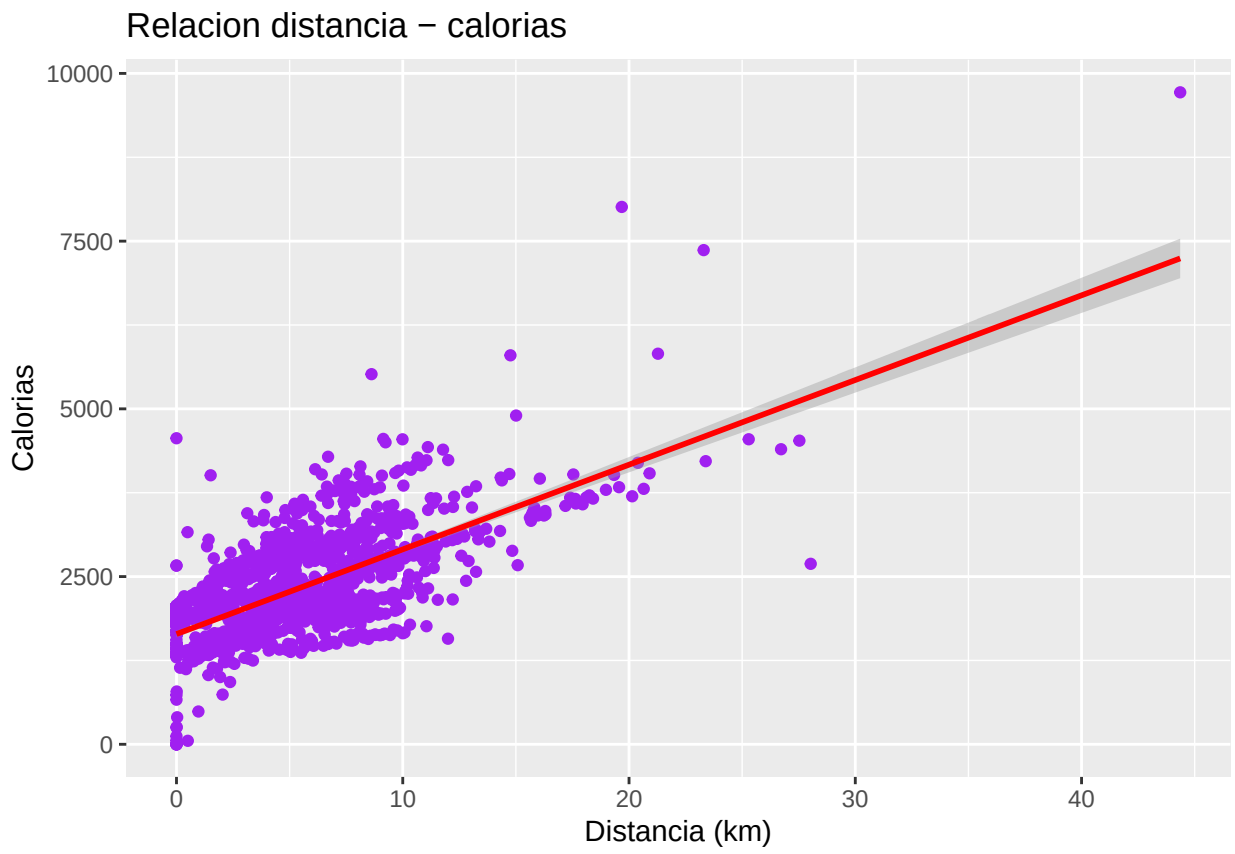


Grafico VIII. Analisis relacion distancia - calorias

```
ggplot (data = master_table_clean, aes (x = TotalDistance, y = Calories)) +
  geom_point (color = "purple") +
  geom_smooth (method = "lm", color = "red") +
  labs (title = "Relacion distancia - calorias",
        x = "Distancia (km)",
        y = "Calorias")
```



Nivel de actividad - dias de la semana

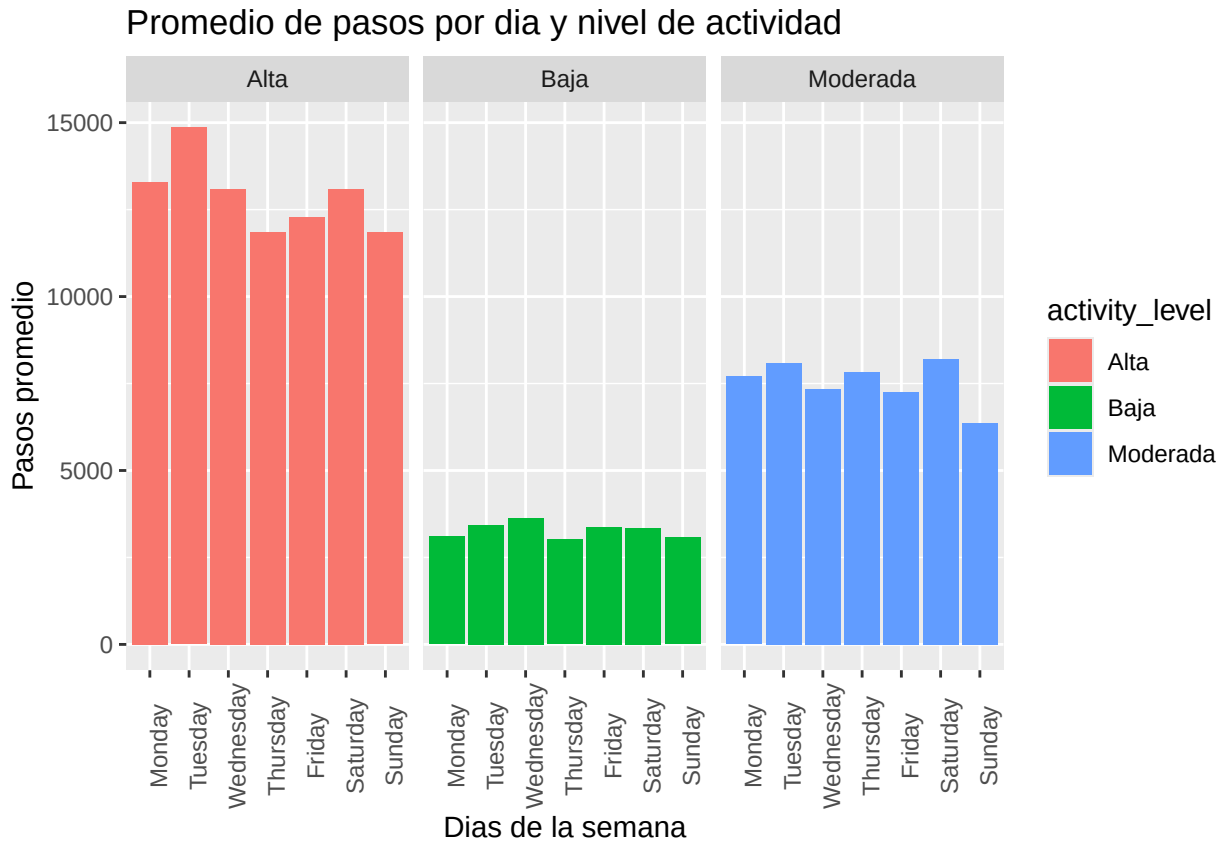
Tabla que distingue dentro de cada dia la cantidad de pasos promedio que realiza cada grupo de actividad

```
activity_weekday <- master_table_clean %>%
  mutate (weekday = weekdays(ActivityDate)) %>%
  left_join (activity_table, by = "Id") %>%
  group_by (activity_level, weekday) %>%
  summarise (avg_steps = mean (TotalSteps), .groups = "drop" ) %>%
  mutate (weekday = factor (weekday,
                            levels = c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturda
  arrange (weekday)
```

Grafico IX. Promedio de pasos al dia segun nivel de actividad

```
ggplot (data = activity_weekday, aes (x = weekday, y = avg_steps, fill = activity_level)) +
  geom_col () +
  facet_wrap (~ activity_level) +
  theme (axis.text.x = element_text(angle = 90)) +
  labs (title = "Promedio de pasos por dia y nivel de actividad",
```

```
        x = "Dias de la semana",
        y = "Pasos promedio")
```



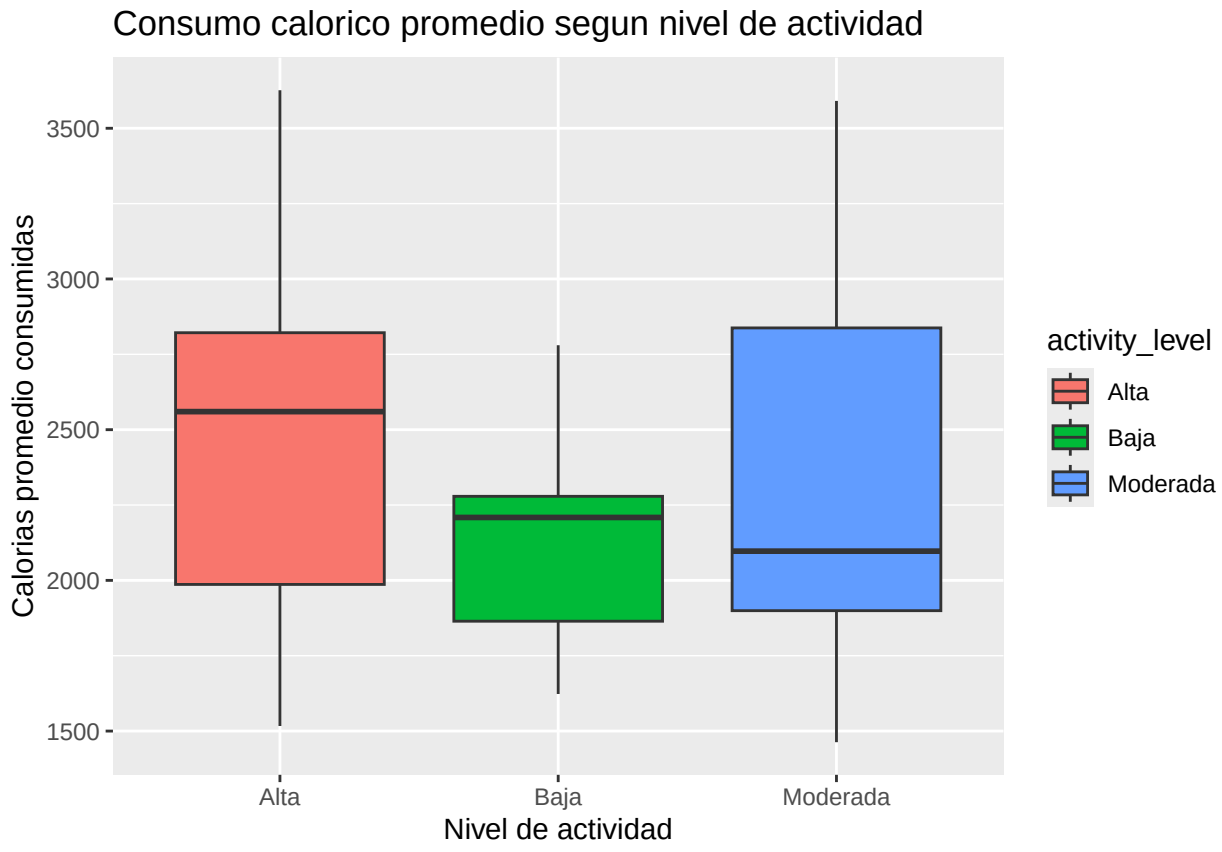
Activity_level vs Calories

Tabla de la relacion entre nivel de actividad y calorias gastadas

```
calories_activity <- master_table_clean %>%
  group_by(Id) %>%
  summarise(
    mean_steps = mean (TotalSteps),
    mean_calories = mean (Calories)) %>%
  mutate (activity_level = case_when (
    mean_steps < 5000 ~ "Baja",
    mean_steps < 10000 ~ "Moderada",
    TRUE ~ "Alta"
  ))
```

Grafico X. Consumo calorico promedio segun nivel de actividad

```
ggplot(data = calories_activity, aes (x = activity_level, y = mean_calories, fill = activity_level)) +
  geom_boxplot() +
  labs (title = "Consumo calorico promedio segun nivel de actividad",
        x = "Nivel de actividad",
        y = "Calorias promedio consumidas")
```



Reconocimiento de perfiles extremos (5 mas y menos activos)

Pasos y calorias por usuario

```
user_calories_activity <- master_table_clean %>%
  group_by(Id) %>%
  summarise(
    mean_steps = mean (TotalSteps),
    mean_calories = mean (Calories),
    active_days = n ())
```

Top 5 mas activos

```
top5 <- user_calories_activity %>%
  arrange (desc(mean_steps)) %>%
```

```
  slice (1:5) %>%
  mutate (group = "Top 5 activos")
```

Top 5 menos activos

```
bottom5 <- user_calories_activity %>%
  arrange (mean_steps) %>%
  slice (1:5) %>%
  mutate (group = "Top 5 menos activos")
```

Grafico XI. Comparacion de perfiles extremos

```
extreme_users <- bind_rows(top5, bottom5)

ggplot (extreme_users, aes (x = reorder(Id, mean_steps), y = mean_calories, fill = group)) +
  geom_col() +
  theme (axis.text.x = element_text(angle = 90)) +
  labs (title = "Top 5 vs Bottom 5 usuarios",
        x = "Usuario",
        y = "Calorias promedio")
```