Google-Data-Analytic-Project-Capstone.R

Dell

2023-05-30

library(readxl)  
activity <- read\_excel("Fitabase Data 4.12.16-5.12.16/dailyActivity\_merged.xlsx") #activity contains daily data on steps, distance, calories, and active minute  
library(readr)  
heartrate <- read\_csv("Fitabase Data 4.12.16-5.12.16/heartrate\_seconds\_merged.csv")

## Rows: 2483658 Columns: 3  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): Time  
## dbl (2): Id, Value  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

weight <- read\_csv("Fitabase Data 4.12.16-5.12.16/weightLogInfo\_merged.csv") #weight contains data on weight in Kg or Pounds

## Rows: 67 Columns: 8  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): Date  
## dbl (6): Id, WeightKg, WeightPounds, Fat, BMI, LogId  
## lgl (1): IsManualReport  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

sleep <- read\_csv("Fitabase Data 4.12.16-5.12.16/sleepDay\_merged.csv") #sleep contains data on the total time spent on bed and the total time asleep each day

## Rows: 413 Columns: 5  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): SleepDay  
## dbl (4): Id, TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

METs <- read\_csv("Fitabase Data 4.12.16-5.12.16/minuteMETsNarrow\_merged.csv") #METs is the metabolic equivalent of task for every minute

## Rows: 1325580 Columns: 3  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): ActivityMinute  
## dbl (2): Id, METs  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

library(pacman) #package used for managing other packages  
library(tidyverse) #package used for data cleaning and manipulation

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ purrr 1.0.1  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

glimpse(activity) #To see the entire columns in activity table created

## Rows: 940  
## Columns: 15  
## $ Id <dbl> 1503960366, 1503960366, 1503960366, 150396036…  
## $ ActivityDate <dttm> 2016-04-12, 2016-04-13, 2016-04-14, 2016-04-…  
## $ TotalSteps <dbl> 13162, 10735, 10460, 9762, 12669, 9705, 13019…  
## $ TotalDistance <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8…  
## $ TrackerDistance <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8…  
## $ LoggedActivitiesDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
## $ VeryActiveDistance <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3.5…  
## $ ModeratelyActiveDistance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1.3…  
## $ LightActiveDistance <dbl> 6.06, 4.71, 3.91, 2.83, 5.04, 2.51, 4.71, 5.0…  
## $ SedentaryActiveDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
## $ VeryActiveMinutes <dbl> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66, 4…  
## $ FairlyActiveMinutes <dbl> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, 21…  
## $ LightlyActiveMinutes <dbl> 328, 217, 181, 209, 221, 164, 233, 264, 205, …  
## $ SedentaryMinutes <dbl> 728, 776, 1218, 726, 773, 539, 1149, 775, 818…  
## $ Calories <dbl> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 203…

## Filter and clean the activity data  
#selecting only the possible columns the will be used for data visualization   
activity\_1 <- activity %>%   
 mutate(date = as\_date(ActivityDate, "%Y-%m-%d")) %>%   
 mutate(month\_year = format(date, "%b %Y")) %>%   
 select(Id, date, ActivityDate, TotalSteps, TotalDistance, VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryMinutes, Calories, month\_year) %>%   
 select(-ActivityDate) %>%   
 filter(!(Id %in% 4057192912))

## Warning: There was 1 warning in `mutate()`.  
## ℹ In argument: `date = as\_date(ActivityDate, "%Y-%m-%d")`.  
## Caused by warning in `as.POSIXlt.POSIXct()`:  
## ! unknown timezone '%Y-%m-%d'

head(activity\_1)

## # A tibble: 6 × 10  
## Id date TotalSteps TotalDistance VeryActiveMinutes  
## <dbl> <date> <dbl> <dbl> <dbl>  
## 1 1503960366 2016-04-12 13162 8.5 25  
## 2 1503960366 2016-04-13 10735 6.97 21  
## 3 1503960366 2016-04-14 10460 6.74 30  
## 4 1503960366 2016-04-15 9762 6.28 29  
## 5 1503960366 2016-04-16 12669 8.16 36  
## 6 1503960366 2016-04-17 9705 6.48 38  
## # ℹ 5 more variables: FairlyActiveMinutes <dbl>, LightlyActiveMinutes <dbl>,  
## # SedentaryMinutes <dbl>, Calories <dbl>, month\_year <chr>

glimpse(METs) #To see the entire columns in METs table created

## Rows: 1,325,580  
## Columns: 3  
## $ Id <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 1503960…  
## $ ActivityMinute <chr> "4/12/2016 12:00:00 AM", "4/12/2016 12:01:00 AM", "4/12…  
## $ METs <dbl> 10, 10, 10, 10, 10, 12, 12, 12, 12, 12, 12, 12, 10, 10,…

METs\_1 <- METs %>%   
 mutate(date\_time = parse\_date\_time(ActivityMinute, "%m/%d/%Y %I:%M:%S %p")) %>%   
 select(-ActivityMinute)  
head(METs\_1)

## # A tibble: 6 × 3  
## Id METs date\_time   
## <dbl> <dbl> <dttm>   
## 1 1503960366 10 2016-04-12 00:00:00  
## 2 1503960366 10 2016-04-12 00:01:00  
## 3 1503960366 10 2016-04-12 00:02:00  
## 4 1503960366 10 2016-04-12 00:03:00  
## 5 1503960366 10 2016-04-12 00:04:00  
## 6 1503960366 12 2016-04-12 00:05:00

glimpse(weight) #To see the entire columns in weight table created

## Rows: 67  
## Columns: 8  
## $ Id <dbl> 1503960366, 1503960366, 1927972279, 2873212765, 2873212…  
## $ Date <chr> "5/2/2016 23:59", "5/3/2016 23:59", "4/13/2016 1:08", "…  
## $ 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> 1.46223e+12, 1.46232e+12, 1.46051e+12, 1.46128e+12, 1.4…

weight\_1 <- weight %>%   
 mutate(ID = as.character(Id)) %>%   
 select(ID, WeightKg) %>%   
 group\_by(ID) %>%   
 drop\_na() %>%   
 summarise(avg\_weight = mean(WeightKg))  
head(weight\_1)

## # A tibble: 6 × 2  
## ID avg\_weight  
## <chr> <dbl>  
## 1 1503960366 52.6  
## 2 1927972279 134.   
## 3 2873212765 57   
## 4 4319703577 72.4  
## 5 4558609924 69.6  
## 6 5577150313 90.7

glimpse(sleep)

## Rows: 413  
## Columns: 5  
## $ Id <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 150…  
## $ SleepDay <chr> "4/12/2016 12:00:00 AM", "4/13/2016 12:00:00 AM", "…  
## $ TotalSleepRecords <dbl> 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …  
## $ TotalMinutesAsleep <dbl> 327, 384, 412, 340, 700, 304, 360, 325, 361, 430, 2…  
## $ TotalTimeInBed <dbl> 346, 407, 442, 367, 712, 320, 377, 364, 384, 449, 3…

sleep\_1 <- sleep %>%   
 mutate(Date\_1 = as.character(SleepDay)) %>%   
 mutate(PST = TotalTimeInBed - TotalMinutesAsleep) %>%   
 separate(Date\_1, c("date", "time"), sep = " ") %>%   
 mutate(Date = as.Date(date, "%m/%d/%Y", optional = FALSE)) %>%   
 select(-TotalSleepRecords, -SleepDay, -date, -time) %>%   
 drop\_na() %>%   
 filter(!(Id %in% c(1644430081, 1844505072, 1927972279, 4020332650, 2320127002, 4558609924, 6775888955, 7007744171, 8053475328)))

## Warning: Expected 2 pieces. Additional pieces discarded in 413 rows [1, 2, 3, 4, 5, 6,  
## 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].

head(sleep\_1)

## # A tibble: 6 × 5  
## Id TotalMinutesAsleep TotalTimeInBed PST Date   
## <dbl> <dbl> <dbl> <dbl> <date>   
## 1 1503960366 327 346 19 2016-04-12  
## 2 1503960366 384 407 23 2016-04-13  
## 3 1503960366 412 442 30 2016-04-15  
## 4 1503960366 340 367 27 2016-04-16  
## 5 1503960366 700 712 12 2016-04-17  
## 6 1503960366 304 320 16 2016-04-19

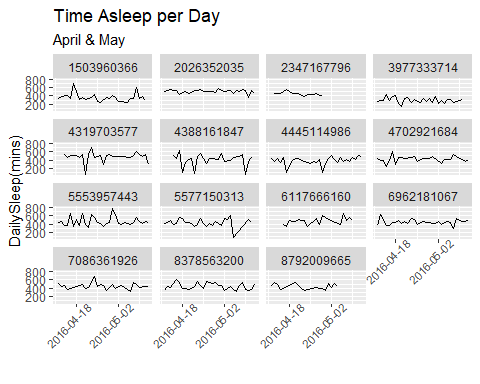
glimpse(heartrate)

## Rows: 2,483,658  
## Columns: 3  
## $ Id <dbl> 2022484408, 2022484408, 2022484408, 2022484408, 2022484408, 2022…  
## $ Time <chr> "4/12/2016 7:21:00 AM", "4/12/2016 7:21:05 AM", "4/12/2016 7:21:…  
## $ Value <dbl> 97, 102, 105, 103, 101, 95, 91, 93, 94, 93, 92, 89, 83, 61, 60, …

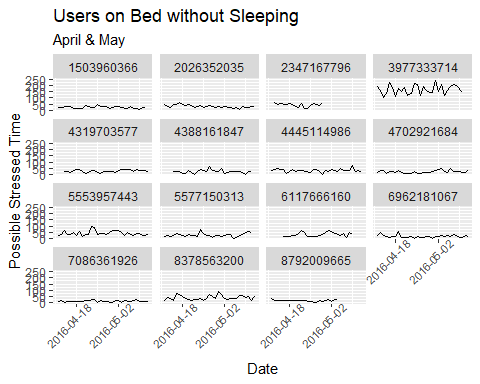
heartrate\_1 <- heartrate %>%   
 mutate(date\_time = parse\_date\_time(Time, "%m/%d/%Y %I:%M:%S %p")) %>%   
 select(-Time)  
head(heartrate\_1)

## # A tibble: 6 × 3  
## Id Value date\_time   
## <dbl> <dbl> <dttm>   
## 1 2022484408 97 2016-04-12 07:21:00  
## 2 2022484408 102 2016-04-12 07:21:05  
## 3 2022484408 105 2016-04-12 07:21:10  
## 4 2022484408 103 2016-04-12 07:21:20  
## 5 2022484408 101 2016-04-12 07:21:25  
## 6 2022484408 95 2016-04-12 07:22:05

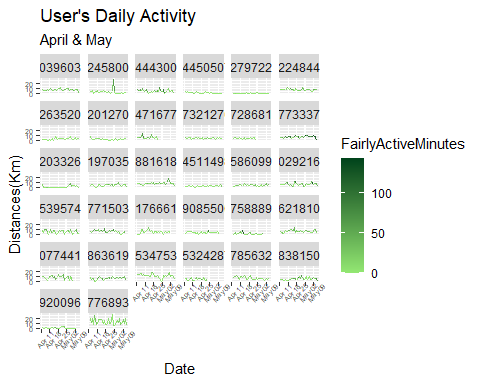
library(ggplot2)  
  
ggplot(sleep\_1,  
 mapping = aes(  
 x = Date,  
 y = TotalMinutesAsleep)  
 )+  
 geom\_line()+  
 labs(x = "",   
 y = "DailySleep(mins)",   
 title = "Time Asleep per Day",   
 subtitle = "April & May")+  
 theme(axis.text.x = element\_text(angle=45,   
 hjust=0.8,   
 size = 8)  
 )+  
 scale\_x\_date(date\_breaks = "2 week")+  
 facet\_wrap(~Id)



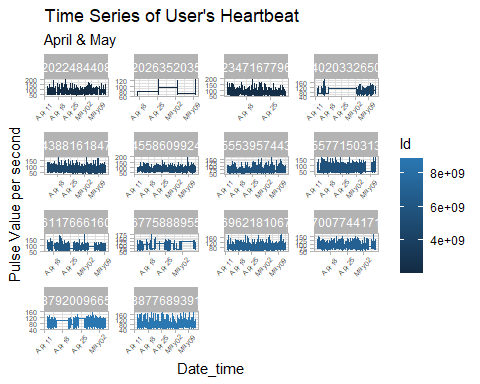
ggplot(sleep\_1,  
 mapping = aes(  
 x = Date,  
 y = PST)  
 )+  
 geom\_line()+  
 labs(x = "Date",   
 y = "Possible Stressed Time",   
 title = "Users on Bed without Sleeping",   
 subtitle = "April & May")+  
 theme(axis.text.x = element\_text(angle=45,   
 hjust=0.8,   
 size = 8)  
 )+  
 scale\_x\_date(date\_breaks = "2 week")+  
 facet\_wrap(~Id)



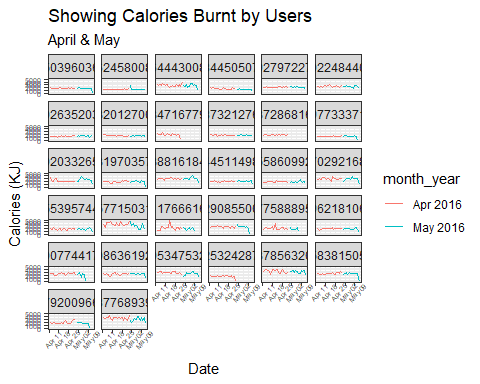
ggplot(activity\_1,  
 mapping = aes(  
 x = date,  
 y = TotalDistance,  
 colour = FairlyActiveMinutes)  
 ) +  
 geom\_line() +  
 scale\_color\_gradient(low = "#94E873", high = "#00441B") +  
 labs(x = "Date",  
 y = "Distances(Km)",  
 title = "User's Daily Activity",  
 subtitle = "April & May"  
 ) +  
 scale\_x\_date(date\_minor\_breaks = "month")+  
 theme(axis.text.x = element\_text(angle=45)  
 )+  
 theme(axis.text.y = element\_text(size = 5))+  
 theme(axis.text.x = element\_text(size = 5))+  
 facet\_wrap(vars(Id))



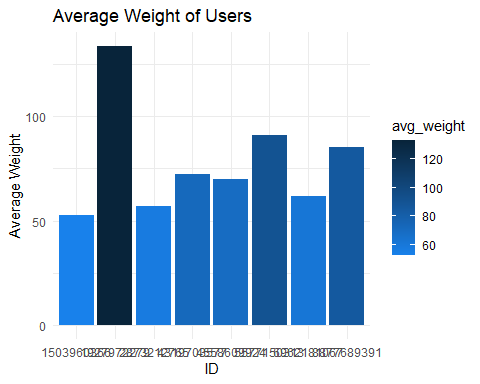
ggplot(heartrate\_1,   
 mapping = aes(  
 x = date\_time,   
 y = Value,   
 colour = Id)  
 ) +  
 geom\_step() +  
 scale\_color\_gradient(low = "#132B43", high = "#2A77B2") +  
 labs(x = "Date\_time",  
 y = "Pulse Value per second",  
 title = "Time Series of User's Heartbeat",  
 subtitle = "April & May"  
 ) +  
 theme\_light() +  
 theme(axis.text.y = element\_text(size = 5))+  
 theme(axis.text.x = element\_text(angle=45,   
 hjust=0.8,  
 size = 5)  
 )+  
 facet\_wrap(vars(Id), scales = "free")



ggplot(activity\_1,   
 mapping = aes(  
 x = date,   
 y = Calories,   
 colour = month\_year)  
 ) +  
 geom\_line() +  
 scale\_color\_hue(direction = 1) +  
 labs(x = "Date",  
 y = "Calories (KJ)",  
 title = "Showing Calories Burnt by Users",  
 subtitle = "April & May"  
 ) +  
 theme\_bw() +  
 theme(axis.text.y = element\_text(size = 5))+  
 theme(axis.text.x = element\_text(angle=45,  
 size = 5)  
 )+  
 facet\_wrap(vars(Id))



ggplot(weight\_1,   
 mapping = aes(  
 x = ID,   
 y = avg\_weight,   
 fill = avg\_weight)  
 )+  
 geom\_col() +  
 theme\_minimal()+  
 labs(x = "ID",  
 y = "Average Weight",  
 title = "Average Weight of Users"  
 ) +  
 scale\_fill\_gradient(low = "#1881EB", high = "#08243A")



print(mean(activity\_1$TotalDistance))

## [1] 5.500929

print(sd(activity\_1$TotalDistance))

## [1] 3.927566

print(mean(sleep\_1$TotalMinutesAsleep)/60)

## [1] 7.143887

print(sd(sleep\_1$TotalMinutesAsleep)/60)

## [1] 1.676945

print(mean(sleep\_1$PST))

## [1] 38.38522

print(sd(sleep\_1$PST))

## [1] 40.85727