Be-Healthy

Pooja Rathee

2022-04-19

First we install required packages and go through them

```
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

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

```
library(ggplot2)
library(purrr)
library(readr)
library(scales)
```

```
##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

#Importing the required data

```
Activity <- read.csv("data/dailyActivity_merged.csv")
Calories<- read.csv("data/dailyCalories_merged.csv")
Intensities <- read.csv("data/dailyIntensities_merged.csv")
Steps <- read.csv("data/dailySteps_merged.csv")
sleepDay <- read.csv("data/sleepDay_merged.csv")
weight <- read.csv("data/weightLogInfo_merged.csv")
```

A quick look in the following columns

```
colnames(Activity)
```

```
##  [1] "Id"                       "ActivityDate"
##  [3] "TotalSteps"               "TotalDistance"
##  [5] "TrackerDistance"          "LoggedActivitiesDistance"
##  [7] "VeryActiveDistance"       "ModeratelyActiveDistance"
##  [9] "LightActiveDistance"      "SedentaryActiveDistance"
## [11] "VeryActiveMinutes"        "FairlyActiveMinutes"
## [13] "LightlyActiveMinutes"     "SedentaryMinutes"
## [15] "Calories"
```

```
colnames(Calories)
```

```
## [1] "Id"          "ActivityDay" "Calories"
```

```
colnames(Intensities)
```

```
##  [1] "Id"                       "ActivityDay"
##  [3] "SedentaryMinutes"         "LightlyActiveMinutes"
##  [5] "FairlyActiveMinutes"      "VeryActiveMinutes"
##  [7] "SedentaryActiveDistance"  "LightActiveDistance"
##  [9] "ModeratelyActiveDistance" "VeryActiveDistance"
```

```
colnames(Steps)
```

```
## [1] "Id"          "ActivityDay" "StepTotal"
```

```
colnames(sleepDay)
```

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

```
colnames(weight)
```

```
## [1] "Id"             "Date"           "WeightKg"       "WeightPounds"
## [5] "Fat"            "BMI"            "IsManualReport" "LogId"
```

Merging activity and sleep data

```
merge_1 <- merge(Activity,Calories, by = c("Id","Calories"))
merge_2 <- merge(Intensities,Intensities, by = c("Id","ActivityDay","SedentaryMinutes", "LightlyActiveM

merge_daily <- merge(merge_1, merge_2, by = c("Id","ActivityDay","SedentaryMinutes", "LightlyActiveMinut
select(-ActivityDay) %>% rename(Date = ActivityDate)

daily_data <- merge(merge_daily, sleepDay, by = "Id",all=TRUE) %>% drop_na() %>% select(-SleepDay, -Tra
```

Take a look on the summary

```
summary(daily_data)
```

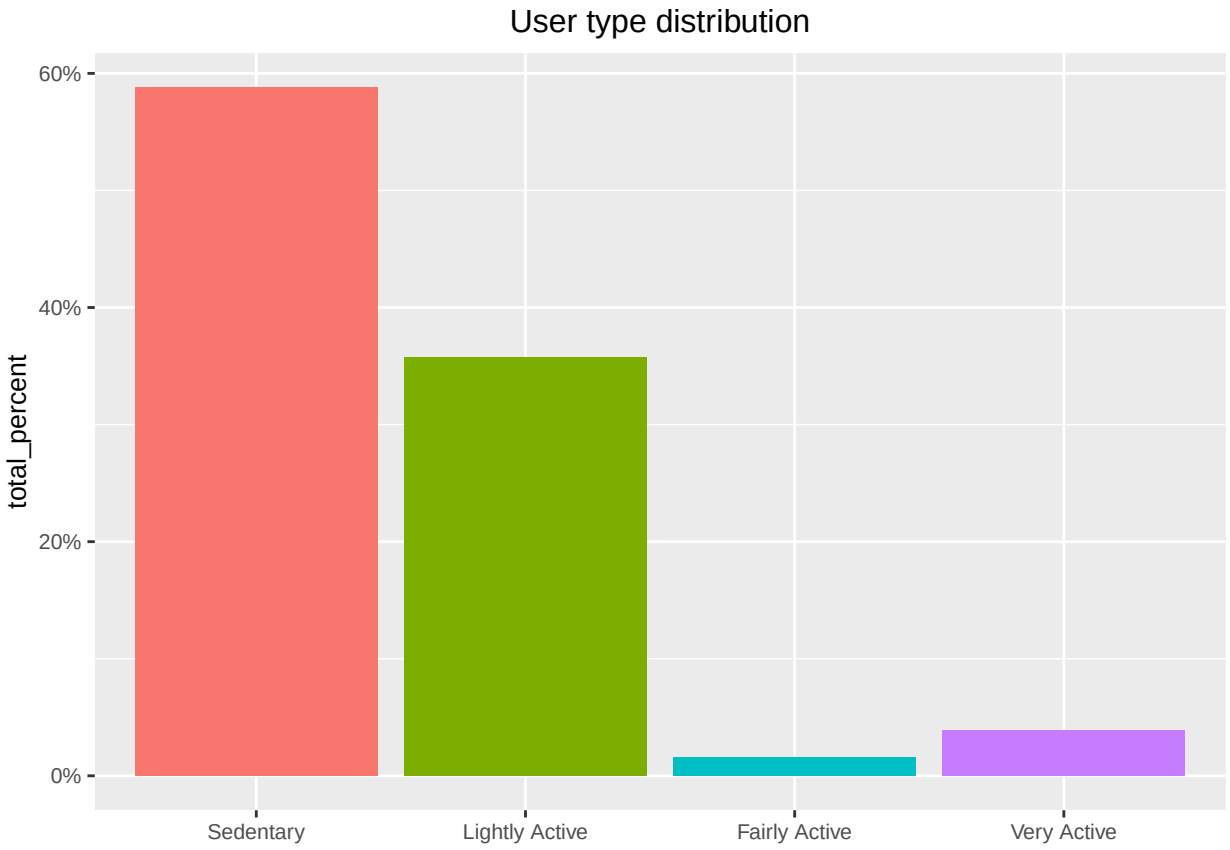
```
##        Id            SedentaryMinutes LightlyActiveMinutes FairlyActiveMinutes
##  Min.   :1.504e+09   Min.   :   0.0   Min.   :  0.0        Min.   :  0.00
##  1st Qu.:4.020e+09   1st Qu.: 687.0   1st Qu.:  0.0        1st Qu.:  0.00
##  Median :4.703e+09   Median : 781.0   Median :171.0        Median :  3.00
##  Mean   :5.117e+09   Mean   : 938.6   Mean   :156.4        Mean   : 13.58
##  3rd Qu.:6.962e+09   3rd Qu.:1440.0   3rd Qu.:240.0        3rd Qu.: 19.00
##  Max.   :8.792e+09   Max.   :1440.0   Max.   :518.0        Max.   :143.00
##  VeryActiveMinutes SedentaryActiveDistance LightActiveDistance
##  Min.   :  0.00    Min.   :0.0000000       Min.   : 0.000
##  1st Qu.:  0.00    1st Qu.:0.0000000       1st Qu.: 0.000
##  Median :  0.00    Median :0.0000000       Median : 2.860
##  Mean   : 18.76    Mean   :0.0005276       Mean   : 2.771
##  3rd Qu.: 28.00    3rd Qu.:0.0000000       3rd Qu.: 4.480
##  Max.   :210.00    Max.   :0.1100000       Max.   :10.300
##  ModeratelyActiveDistance VeryActiveDistance    Calories        Date
##  Min.   :0.0000           Min.   : 0.000     Min.   :   0   Length:15901
##  1st Qu.:0.0000           1st Qu.: 0.000     1st Qu.:1693   Class :character
##  Median :0.1100           Median : 0.000     Median :2013   Mode  :character
##  Mean   :0.5729           Mean   : 1.094     Mean   :2220
##  3rd Qu.:0.7900           3rd Qu.: 1.740     3rd Qu.:2643
##  Max.   :6.4800           Max.   :13.400     Max.   :4900
##    TotalSteps    TotalDistance    LoggedActivitiesDistance TotalSleepRecords
##  Min.   :    0   Min.   : 0.000   Min.   :0.00000          Min.   :1.000
##  1st Qu.:    0   1st Qu.: 0.000   1st Qu.:0.00000          1st Qu.:1.000
##  Median : 6393   Median : 4.480   Median :0.00000          Median :1.000
##  Mean   : 6351   Mean   : 4.487   Mean   :0.09649          Mean   :1.116
##  3rd Qu.:10460   3rd Qu.: 7.390   3rd Qu.:0.00000          3rd Qu.:1.000
##  Max.   :22988   Max.   :17.950   Max.   :4.94214          Max.   :3.000
##  TotalMinutesAsleep TotalTimeInBed
##  Min.   : 58.0      Min.   : 61.0
##  1st Qu.:360.0      1st Qu.:402.0
##  Median :427.0      Median :459.0
##  Mean   :417.3      Mean   :456.1
##  3rd Qu.:490.0      3rd Qu.:530.0
##  Max.   :796.0      Max.   :961.0
```

Categorize the data on the basis of active minutes

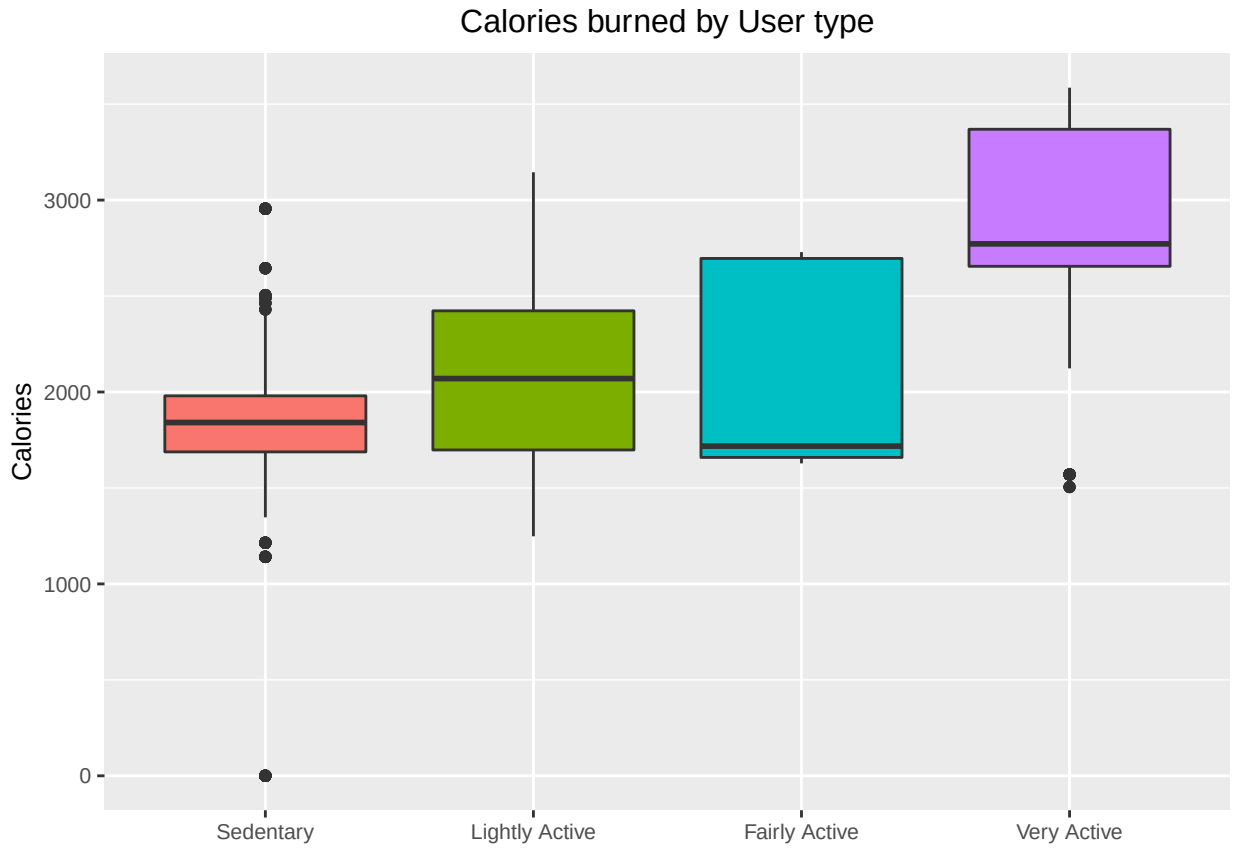
```
data_by_usertype <- daily_data %>%
summarise(
user_type = factor(case_when(
    SedentaryMinutes > mean(SedentaryMinutes) & LightlyActiveMinutes < mean(LightlyActiveMinutes) & Fai
    SedentaryMinutes < mean(SedentaryMinutes) & LightlyActiveMinutes > mean(LightlyActiveMinutes) & Fai
    SedentaryMinutes < mean(SedentaryMinutes) & LightlyActiveMinutes < mean(LightlyActiveMinutes) & Fai
    SedentaryMinutes < mean(SedentaryMinutes) & LightlyActiveMinutes < mean(LightlyActiveMinutes) & Fai
),levels=c("Sedentary", "Lightly Active", "Fairly Active", "Very Active")), Calories, .group=Id) %>% dr
```

Visualize both the user type distribution and the calories burned for every user type:

```
data_by_usertype %>%
group_by(user_type) %>%
summarise(total = n()) %>%
mutate(totals = sum(total)) %>%
group_by(user_type) %>%
summarise(total_percent = total / totals) %>%
ggplot(aes(user_type,y=total_percent, fill=user_type)) +
    geom_col()+
    scale_y_continuous(labels = scales::percent) +
    theme(legend.position="none") +
    labs(title="User type distribution", x=NULL) +
    theme(legend.position="none", text = element_text(size = 10),plot.title = element_text(hjust = 0.5))
```

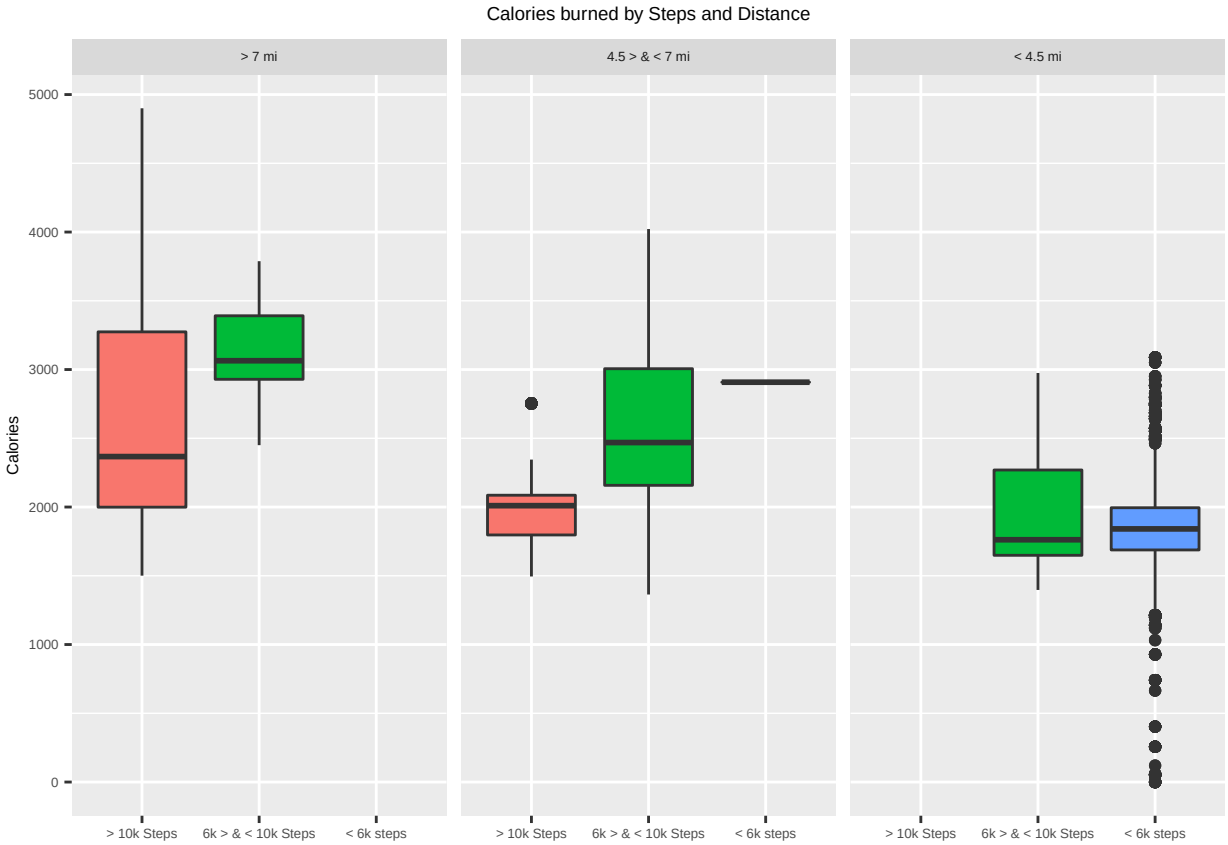



```
ggplot(data_by_usertype, aes(user_type, Calories, fill=user_type)) +
    geom_boxplot() +
    theme(legend.position="none") +
    labs(title="Calories burned by User type", x=NULL) +
    theme(legend.position="none", text = element_text(size = 10),plot.title = element_text(hjust = 0.5))
```



The users are mostly sedentary or Lightly active users but it's very interesting to see that even though they are the biggest category the Fairly active and most importantly, the Very active are the ones with more calories burned. # To check the relation between Distance/Steps and Calories burned by plotting them:

```
daily_data %>%
summarise(
distance = factor(case_when(
    TotalDistance < 4.5 ~ "< 4.5 mi",
    TotalDistance >= 4.5 & TotalDistance <= 7 ~ "4.5 > & < 7 mi",
    TotalDistance > 7 ~ "> 7 mi",
),levels = c("> 7 mi","4.5 > & < 7 mi","< 4.5 mi")),
steps = factor(case_when(
    TotalSteps < 6000 ~ "< 6k steps",
    TotalSteps >= 6000 & TotalSteps <= 10000 ~ "6k > & < 10k Steps",
    TotalSteps > 10000 ~ "> 10k Steps",
),levels = c("> 10k Steps","6k > & < 10k Steps","< 6k steps")),
Calories) %>%
ggplot(aes(steps,Calories,fill=steps)) +
    geom_boxplot() +
    facet_wrap(~distance)+
    labs(title="Calories burned by Steps and Distance",x=NULL) +
    theme(legend.position="none", text = element_text(size = 6),plot.title = element_text(hjust = 0.5))
```



The most calories burned are the "6k > & < 10k Steps" and "> 7 miles" which would indicate some kind of running activity that allows the user to traverse more distance with less steps. # Understanding some summary statistics by visualization

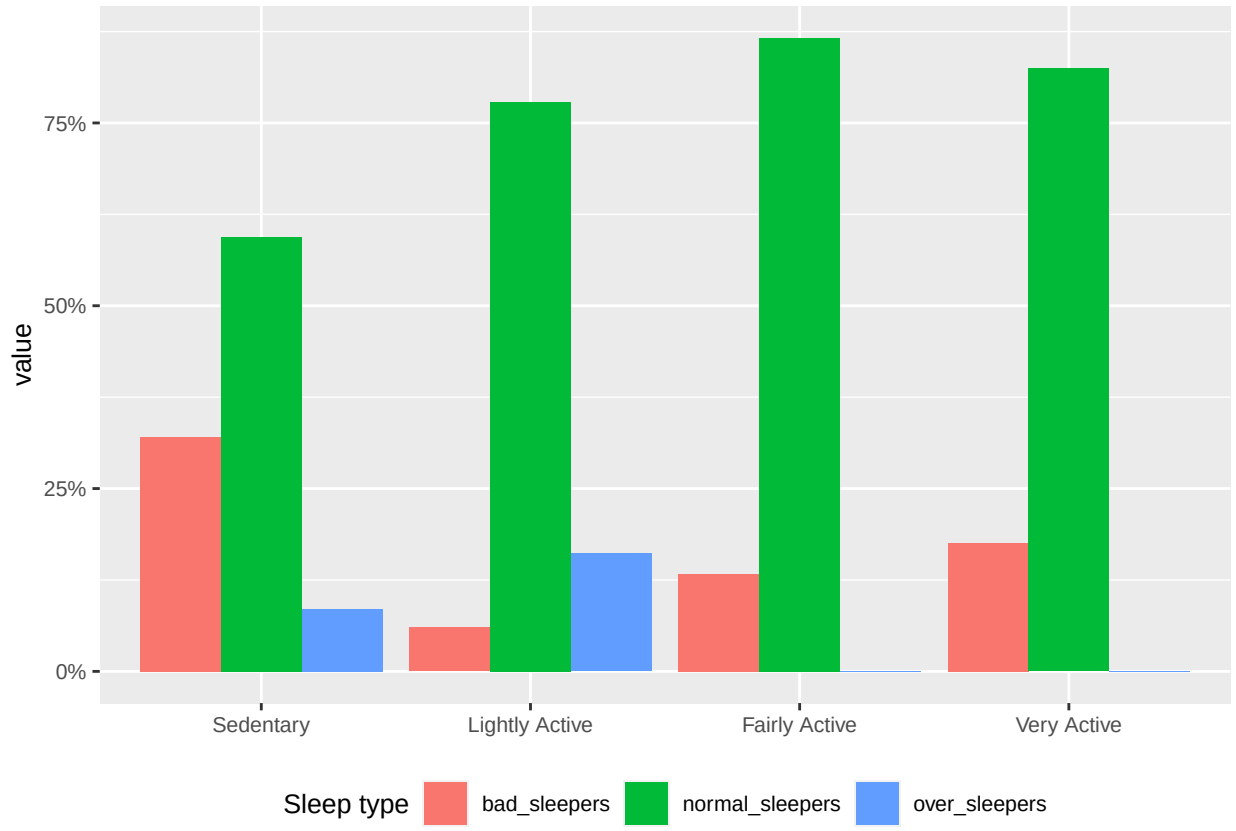
```
#Now let's focus on the sleep quality, for that I will now make categories for the sleeping time and I
sleepType_by_userType <- daily_data %>%
group_by(Id) %>%
summarise(user_type = factor(case_when(
    SedentaryMinutes > mean(SedentaryMinutes) & LightlyActiveMinutes < mean(LightlyActiveMinutes) & Fai
    SedentaryMinutes < mean(SedentaryMinutes) & LightlyActiveMinutes > mean(LightlyActiveMinutes) & Fai
    SedentaryMinutes < mean(SedentaryMinutes) & LightlyActiveMinutes < mean(LightlyActiveMinutes) & Fai
    SedentaryMinutes < mean(SedentaryMinutes) & LightlyActiveMinutes < mean(LightlyActiveMinutes) & Fai
),levels=c("Sedentary", "Lightly Active", "Fairly Active", "Very Active")),
sleep_type = factor(case_when(
    mean(TotalMinutesAsleep) < 360 ~ "Bad Sleep",
    mean(TotalMinutesAsleep) > 360 & mean(TotalMinutesAsleep) <= 480 ~ "Normal Sleep",
    mean(TotalMinutesAsleep) > 480 ~ "Over Sleep"),levels=c("Bad Sleep", "Normal Sleep", "Over Sleep"))
) %>%
drop_na() %>%
group_by(user_type) %>%
summarise(bad_sleepers = sum(sleep_type == "Bad Sleep"), normal_sleepers = sum(sleep_type == "Normal Sl
group_by(user_type) %>%
summarise(
    bad_sleepers = bad_sleepers / total,
    normal_sleepers = normal_sleepers / total,
    over_sleepers = over_sleepers / total,
    .groups="drop"
```

```
)
```

#Now we can plot the data for each user type:

```
sleepType_by_userType_melted<- melt(sleepType_by_userType, id.vars = "user_type")

ggplot(sleepType_by_userType_melted, aes(user_type, value, fill = variable)) +
geom_bar(position = "dodge", stat = "identity") +
scale_y_continuous(labels = scales::percent) +
labs(x=NULL, fill="Sleep type") +
theme(legend.position="bottom",text = element_text(size = 10),plot.title = element_text(hjust = 0.5))
```



It shows very clearly the relation between the activity level and sleep quality as in the sedentary users we find the largest percentage of bad sleepers and with some activity (even very little activity) we see a great increase of normal sleepers.