250401_GoogleCapstone_Bella

HongTran

2025-03-31

```
# Set a CRAN mirror before installing any packages
options(repos = c(CRAN = "https://cloud.r-project.org"))

# Install necessary packages if not already installed
install.packages("data.table")
```

```
## package 'data.table' successfully unpacked and MD5 sums checked

##
## The downloaded binary packages are in
##  C:\Users\mauth\AppData\Local\Temp\RtmpOuAgkN\downloaded_packages
```

```
install.packages("dplyr")
```

```
## package 'dplyr' successfully unpacked and MD5 sums checked

##
## The downloaded binary packages are in
##  C:\Users\mauth\AppData\Local\Temp\RtmpOuAgkN\downloaded_packages
```

```
install.packages("ggplot2")
```

```
## package 'ggplot2' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\mauth\AppData\Local\Temp\RtmpOuAgkN\downloaded_packages
```

```
install.packages("bit64")
```

```
## package 'bit64' successfully unpacked and MD5 sums checked

##
## The downloaded binary packages are in
##  C:\Users\mauth\AppData\Local\Temp\RtmpOuAgkN\downloaded_packages
```

```
install.packages("tinytex")
```

```
## package 'tinytex' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\mauth\AppData\Local\Temp\RtmpOuAgkN\downloaded_packages
```

```
# Now, install TinyTeX distribution

tinytex::install_tinytex(force = TRUE)

# Load libraries
library(data.table)
library(dplyr)
library(ggplot2)
library(bit64)
```

Set Dataset Path & Load Data

```
# Load datasets
daily_activity <- fread("D:/5A. Data_analysis_fundalmental/mturkfitbit_export_4.12.16-5.12.16/Fitabase
sleep_day <- fread("D:/5A. Data_analysis_fundalmental/mturkfitbit_export_4.12.16-5.12.16/Fitabase Data
daily_calories <- fread("D:/5A. Data_analysis_fundalmental/mturkfitbit_export_4.12.16-5.12.16/Fitabase
```

Exploring Data

Column Names

```
colnames(daily_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(daily_calories)
```

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

```
colnames(sleep_day)
```

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

Unique Participants

```
n_distinct(daily_activity$Id)
```

```
## [1] 33
```

```
n_distinct(daily_calories$Id)
```

```
## [1] 33
```

```
n_distinct(sleep_day$Id)
```

```
## [1] 24
```

The number of participants in each attributes are varies ascross the dataset.

Number of Observations

```
nrow(daily_activity)
```

```
## [1] 940
```

```
nrow(daily_calories)
```

```
## [1] 940
```

```
nrow(sleep_day)
```

```
## [1] 413
```

Each characteristic contains a different number of observations.

Summary Statistics

```
library(knitr)
kable(summary(daily_activity[, .(VeryActiveDistance, ModeratelyActiveDistance, LightActiveDistance)]))
```

VeryActiveDistance	ModeratelyActiveDistance	LightActiveDistance
Min. : 0.000	Min. :0.0000	Min. : 0.000
1st Qu.: 0.000	1st Qu.:0.0000	1st Qu.: 1.945
Median : 0.210	Median :0.2400	Median : 3.365
Mean : 1.503	Mean :0.5675	Mean : 3.341

VeryActiveDistance	ModeratelyActiveDistance	LightActiveDistance
3rd Qu.: 2.053	3rd Qu.:0.8000	3rd Qu.: 4.782
Max. :21.920	Max. :6.4800	Max. :10.710

```
kable(summary(daily_activity[, .(VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, Sedentary
```

VeryActiveMinutes	FairlyActiveMinutes	LightlyActiveMinutes	SedentaryMinutes
Min. : 0.00	Min. : 0.00	Min. : 0.0	Min. : 0.0
1st Qu.: 0.00	1st Qu.: 0.00	1st Qu.:127.0	1st Qu.: 729.8
Median : 4.00	Median : 6.00	Median :199.0	Median :1057.5
Mean : 21.16	Mean : 13.56	Mean :192.8	Mean : 991.2
3rd Qu.: 32.00	3rd Qu.: 19.00	3rd Qu.:264.0	3rd Qu.:1229.5
Max. :210.00	Max. :143.00	Max. :518.0	Max. :1440.0

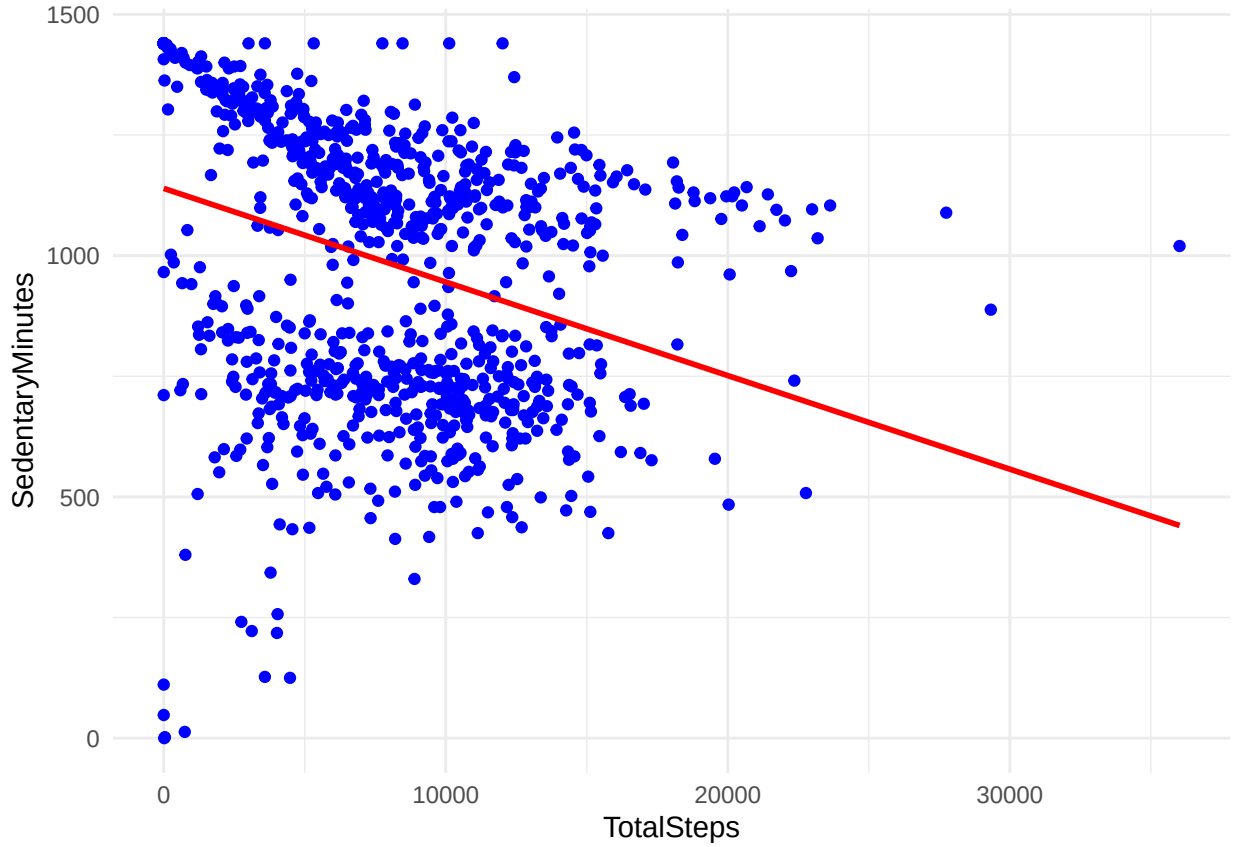
```
**Active distance**

-Across all activity levels, the minimum distance recorded is zero, indicating instances where the acti

-The maximum distances vary, with very active distance having the highest maximum, followed by light ac

-The average distances show a clear trend: light activity contributes the most to overall distance, foll

**Active minutes**

-There appears to be a general trend: as the intensity of activity increases (from light to moderate to

-The distributions for very and moderately active distances and minutes seem to be right-skewed (mean >

-Light activity shows a more symmetrical distribution for both distance and minutes.

-Sedentary time is significantly higher than any of the active time categories.
```

Data Visualization

Steps vs. Sedentary Minutes

```
ggplot(daily_activity, aes(x=TotalSteps, y=SedentaryMinutes)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



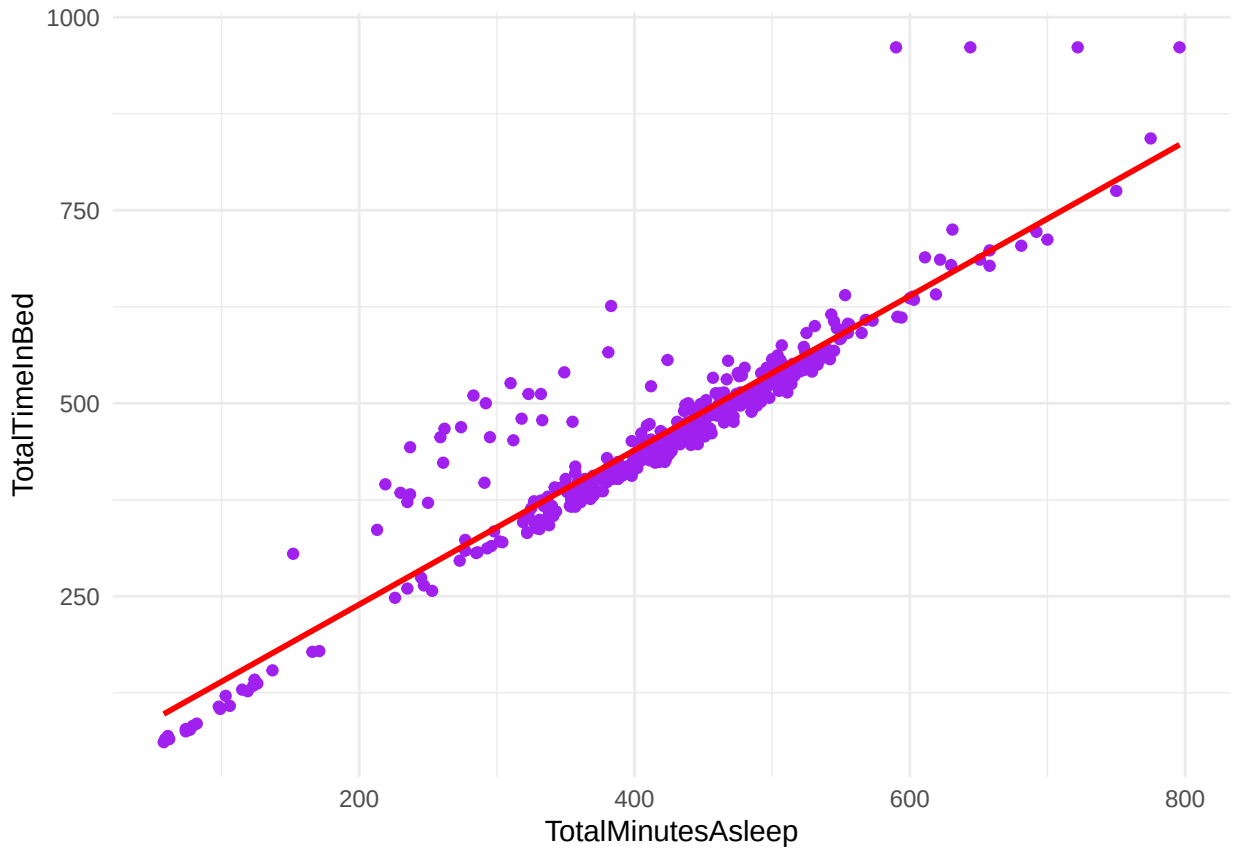
Negative Correlation:

There appears to be a general negative trend between the number of total steps and sedentary minutes. As

Sleep Minutes vs. Time in Bed

```
ggplot(sleep_day, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) +
  geom_point(color = "purple") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Strong Positive Correlation:

There's a very clear positive correlation between the total minutes asleep and the total time spent in

Merging Datasets

```
library(dplyr)
combined_data_1 <- inner_join(daily_calories, daily_activity, by = "Id")
```

```
## Warning in inner_join(daily_calories, daily_activity, by = "Id"): Detected an unexpected many-to-many
## i Row 1 of `x` matches multiple rows in `y`.
## i Row 1 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

```
combined_data_2 <- inner_join(sleep_day, daily_activity, by = "Id")
```

```
## Warning in inner_join(sleep_day, daily_activity, by = "Id"): Detected an unexpected many-to-many rela
## i Row 1 of `x` matches multiple rows in `y`.
## i Row 1 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

Number of Participants in Combined Datasets

```
n_distinct(combined_data_1$Id)
```

```
## [1] 33
```

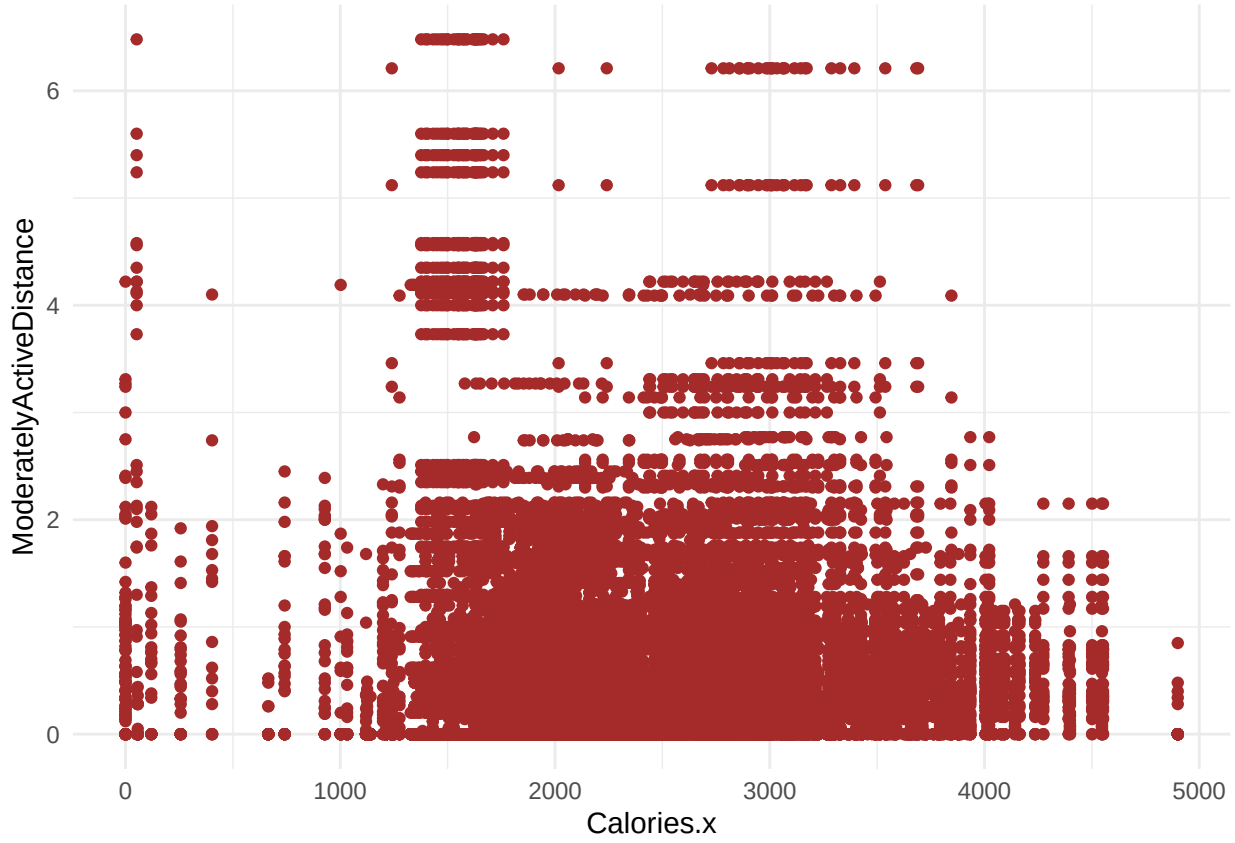
```
n_distinct(combined_data_2$Id)
```

```
## [1] 24
```

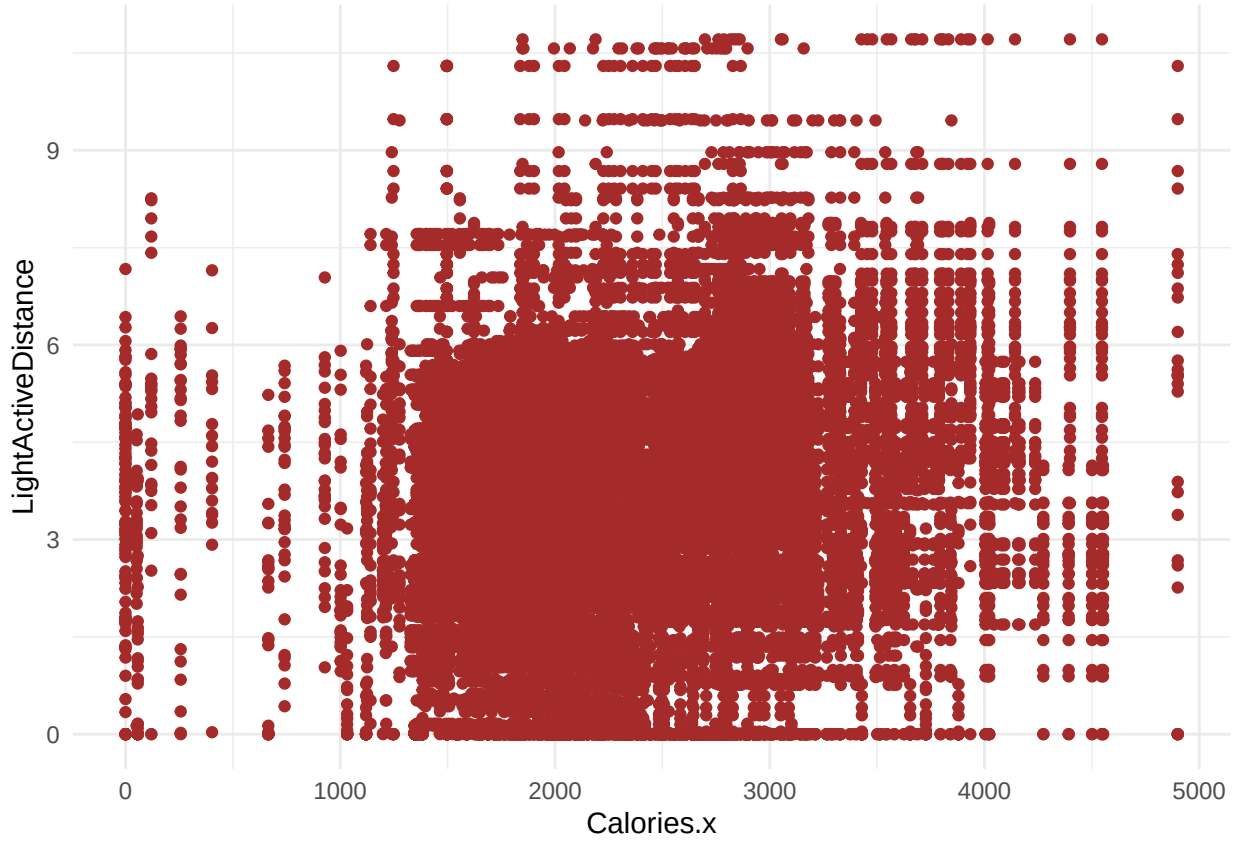
Activity vs. Calories Burned

```
ggplot(combined_data_1, aes(x = Calories.x, y = VeryActiveDistance)) +
  geom_point(color = "brown") +
  theme_minimal()
```

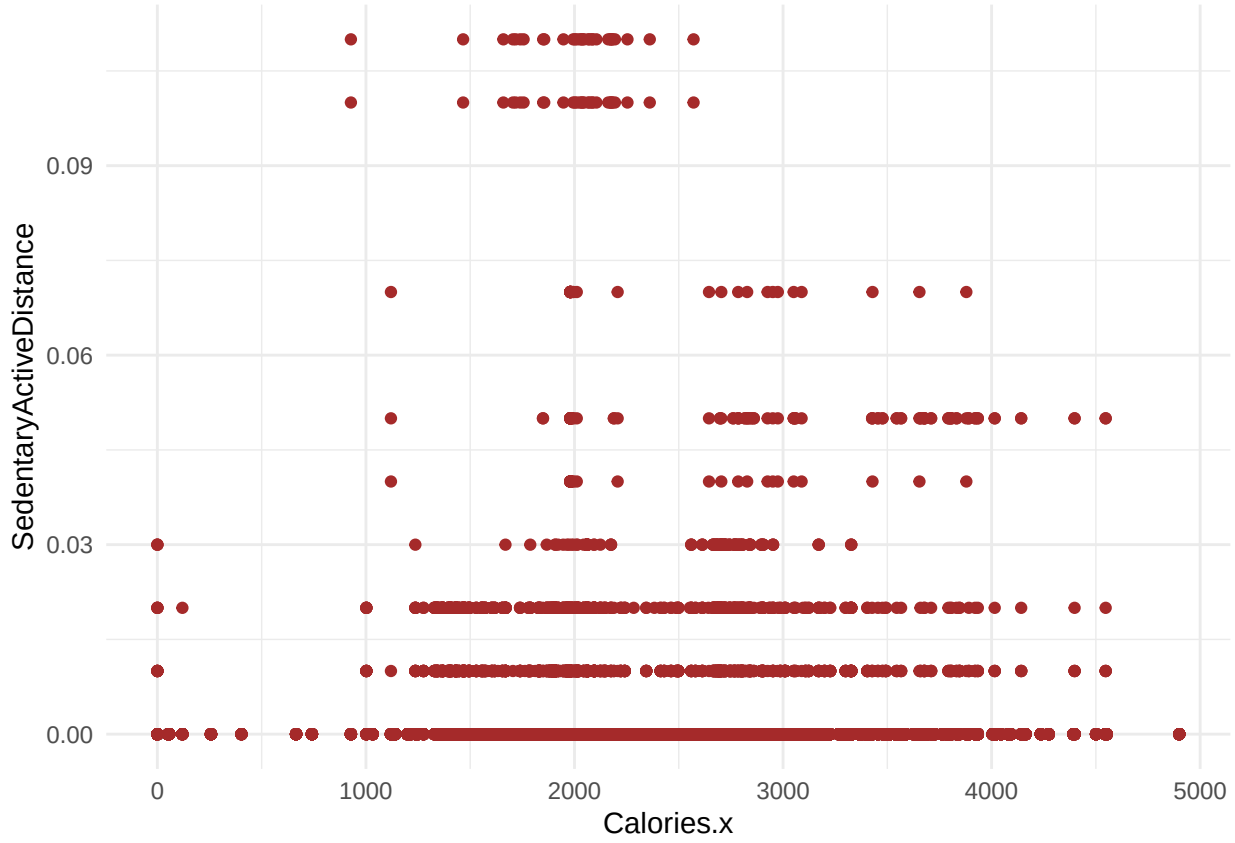

```
ggplot(combined_data_1, aes(x = Calories.x, y = ModeratelyActiveDistance)) +
  geom_point(color = "brown") +
  theme_minimal()
```




```
ggplot(combined_data_1, aes(x = Calories.x, y = LightActiveDistance)) +
  geom_point(color = "brown") +
  theme_minimal()
```




```
ggplot(combined_data_1, aes(x = Calories.x, y = SedentaryActiveDistance)) +
  geom_point(color = "brown") +
  theme_minimal()
```



Across all levels of activity distance (very, moderately, and light), there isn't a strong positive lin

The majority of the recorded distances for the more intense activities (very and moderately active) ten

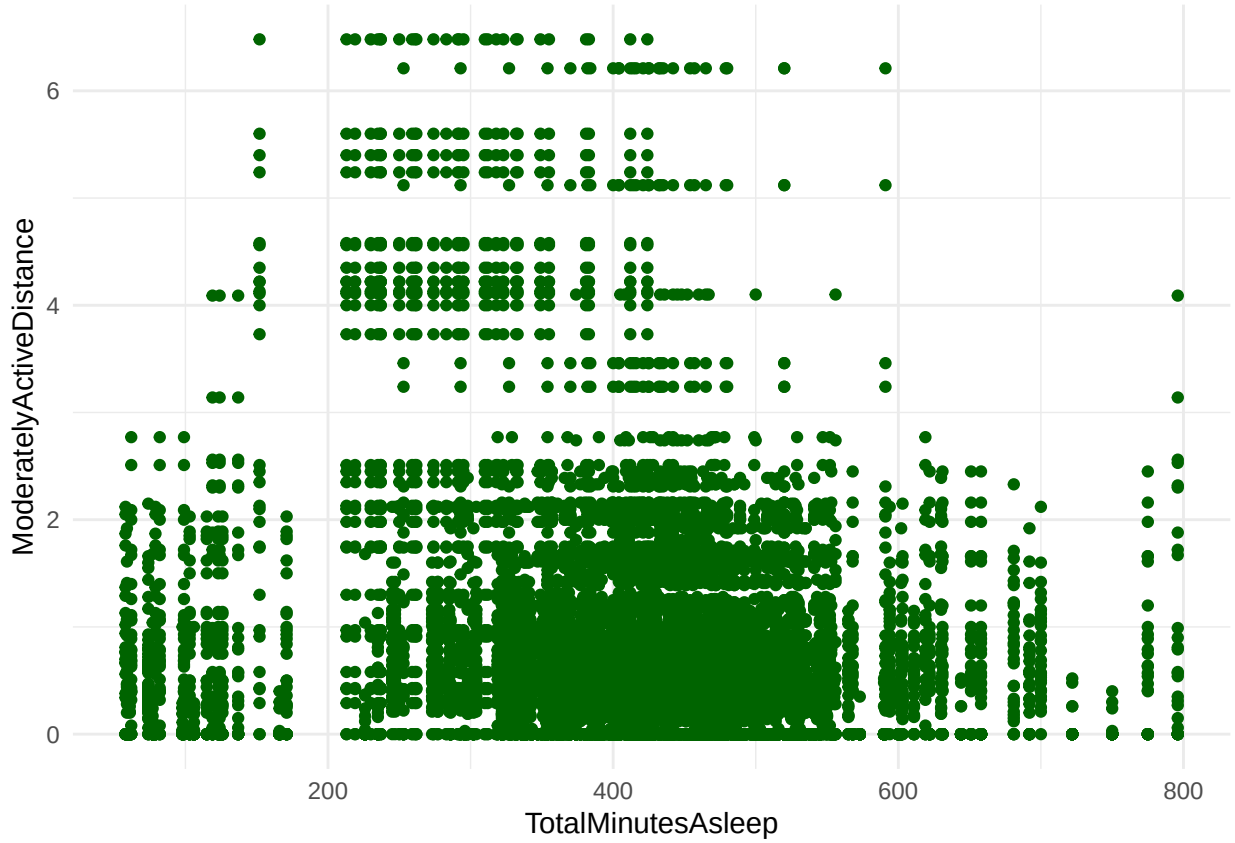
Light active distance shows a broader range of values, indicating that people accumulate more distance

Sedentary active distance is negligible, as expected.

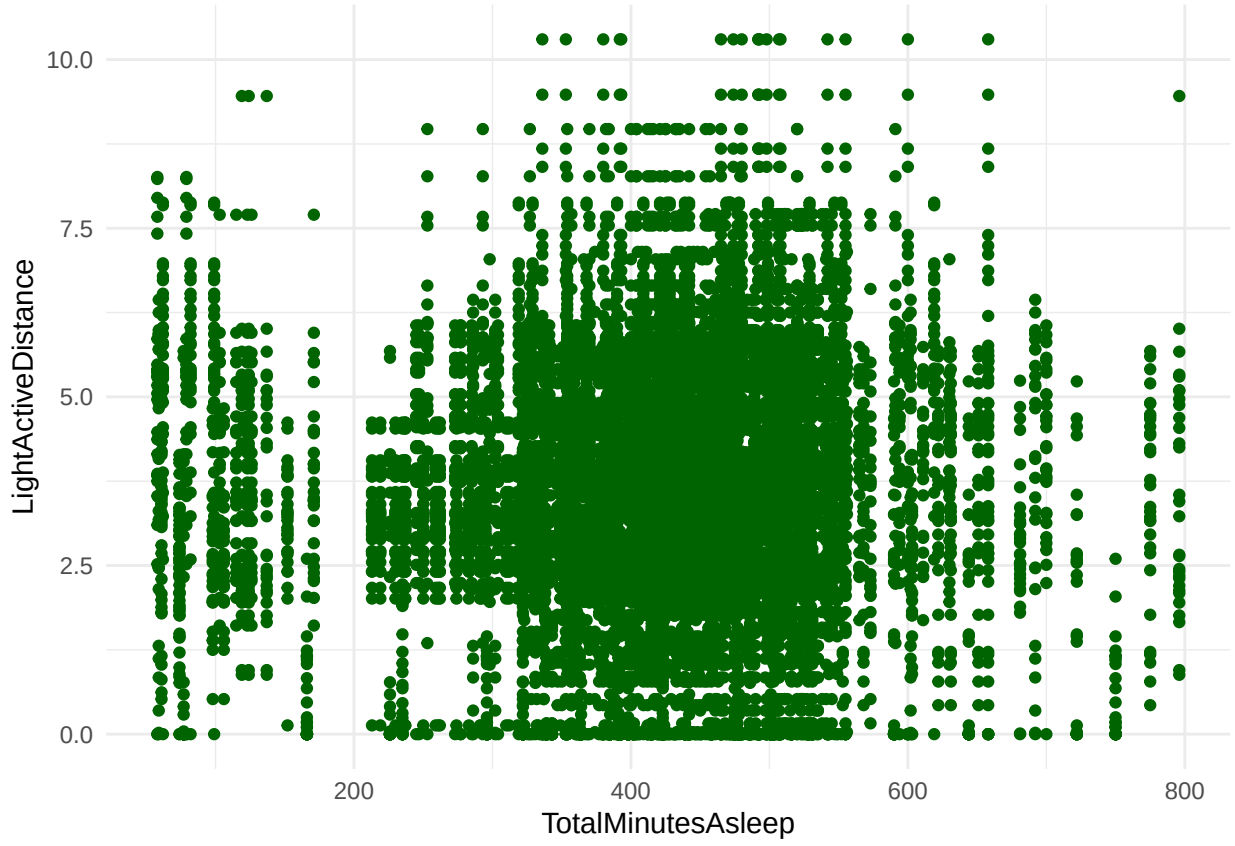
Activity vs. Sleep Minutes

```
ggplot(combined_data_2, aes(x = TotalMinutesAsleep, y = VeryActiveDistance)) +
  geom_point(color = "darkgreen") +
  theme_minimal()
```

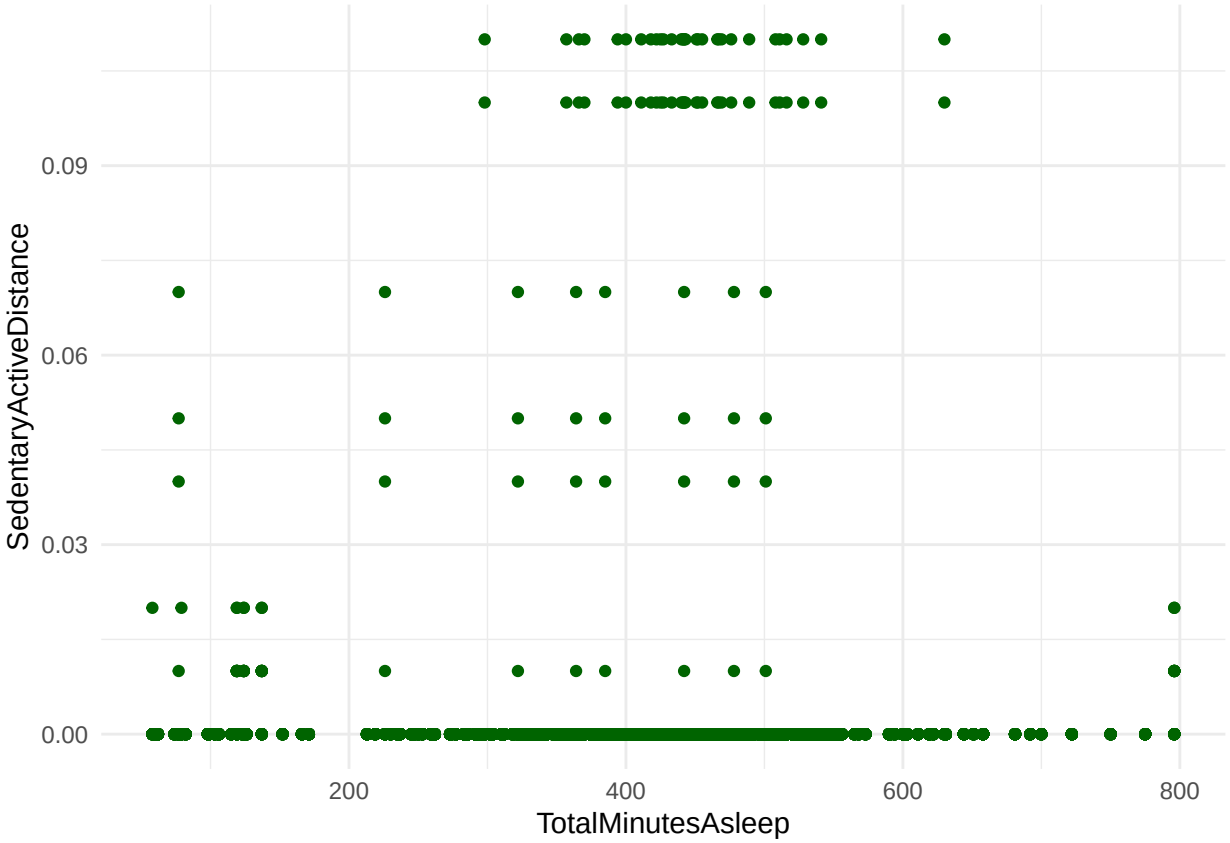

```
ggplot(combined_data_2, aes(x = TotalMinutesAsleep, y = ModeratelyActiveDistance)) +
  geom_point(color = "darkgreen") +
  theme_minimal()
```




```
ggplot(combined_data_2, aes(x = TotalMinutesAsleep, y = LightActiveDistance)) +
  geom_point(color = "darkgreen") +
  theme_minimal()
```




```
ggplot(combined_data_2, aes(x = TotalMinutesAsleep, y = SedentaryActiveDistance)) +
  geom_point(color = "darkgreen") +
  theme_minimal()
```



None of these plots show a strong linear relationship between the total minutes asleep and the distance

The amount of sleep an individual gets doesn't seem to be a primary predictor of how much distance they

This suggests that the factors influencing activity distances are likely more related to daily routines

The consistently low values for SedentaryActiveDistance are expected, regardless of sleep duration.