Fitbit and Exercise Data Study – Report

Steven Gilandas

Table of Contents

Problem Statement, Questions and Objectives 2
The Data: 2
Dataset 1 – Fitbit 2
Dataset 2 – Exercise lab 3
1. Data acquisition and cleaning 4
Participants by metric 4
Data cleaning summary 5
2. EDA 5
Relationship of TotalSteps → Calories burned (across all participants) 6
Relationship between minutes spent at different ActivityLevels and Daily Calories Burned 6
Active minutes vs active distance 8
How did participants divide their time between the 4 activity levels? 9
Distribution of Average Steps Per Day across all Participants 10
Correlation matrix – Activity and Sleep Metrics 10
Pairs plot using R: Activity and Sleep 12
How frequently were participants logging their daily activity? 13
How did activity vary across days of the week? 16
Weekday vs weekend 17
Activity across hours of the day 18
3. Preparing explanatory and response variables, and regression analysis 19
3A. Estimate BMR for each participant by regressing TotalSteps against Calories 19
3B. Group mean calories burned per day by participant Id 22
3C. Look for further correlations between metrics 23
3D. Exercise Lab Dataset 25
Cleaning, dataset exploration, EDA 25
3E. Regression analysis 28
Dummy variable encoding for Gender category 28
(Model 1) Linear model with limited features Calories ~ Duration, Heart_Rate, Body_Temp 29
(Model 2) – Linear model with all features Calories ~ Duration, Heart_Rate, Body_Temp, Height, Weight, Age, Gender 30
(Model 3) – Polynomial model with all features, and interaction between main predictors Calories ~ Duration, Heart_Rate, Body_Temp, Height, Weight, Age, Gender 31
(Model 4) – Simplified linear regression with limited interaction Calories ~(Heart_rate)^3, (Duration)^2, (Body_Temp)^2, Height, Weight, Age, Gender, Height*Weight 33
4. General conclusions and recommendation 34
General remarks: 34
Dataset 1: Fitbit 34
Dataset 2: Exercise Lab 36
Business Recommendations: 36
Limitations and Future Data Collection Strategies 37
Appendix A: R code for Pairs Plot 38

Introduction

This study involves looking at 2 separate datasets:

1) Activity/ sleep metrics for 33 participants over 31 days measured using a Fitbit wearable device.
2) Exercise lab data for an individual exercise session: calories burned during the session was recorded along with activity and demographic information for each participant.

Problem Statement, Questions and Objectives

Dataset 1: Month activity and sleep - Fitbit

- What is the average Fitbit user's pattern of activity and sleep – looking at daily and weekly timeframes?
- Can we find a relationship between users' pattern of activity and daily calorie burn/ Basal Metabolic Rate?
- How regularly do users log their activity over the month?
    - Is there a relationship between frequency of logged activity and daily calorie burn, or other health metrics?

Dataset 2: Individual exercise session – Exercise lab

- For an individual exercise session, what are the most important metrics that can be used to predict calorie burn?
- Using metrics that could normally be obtained from a Fitbit wearable, what is the most accurate model we can build to predict Calorie burn for the exercise session?

General questions

- Can any insights be gained into app or hardware features that Fitbit might be able to incorporate in future iterations of their technology?
- What are the minimum metrics that Fitbit needs to obtain to calculate the daily calorie burn of users (demographic information + metrics obtained from wearable)?
- Does regular logging of activity correlate to improved fitness?

The Data:

Dataset 1 – Fitbit

- Public domain data obtained from Kaggle (uploaded by user Mobius):
    - https://www.kaggle.com/datasets/arashnic/fitbit

    - 18 .csv files
- Feature columns we will consider:
    - **Id** – Unique participant id
    - **ActivityDate** – Date of activity recorded (for daily activities)
    - **ActivityHour** – Datetime for hourly granularity
    - **Measures of distance:**
        - TotalSteps, TotalDistance
        - Distance broken down by activity intensity:
            - VeryActiveDistance, FairlyActiveDistance, LightlyActiveDistance, SedentaryDistance
    - **Time breakdown of activity:**
        - VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryMinutes
    - **Calories** – The calories burned during the time interval considered (day/hour)
- **Note:** the Fitbit app uses distance metrics (steps, distance travelled, distance travelled by activity level) to calculate Calories burned. Hence, these distance metrics will be autocorrelated with Calories. Intensity level of activity is also identified by a calculation of distance travelled/ time. Consequently, time spent at different activity intensity levels (VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedetaryMinutes) also has an indirect mathematical relationship with Calories as seen in the Fitbit app. Given these mathematical relationships, we should be careful of including these metrics as predictors of Calories in any model that we create.
- Data was collected from **33 participants** over a period of **31 days** (12.04.2016-12.05.2016). Participants had to opt in to having their information collected via a survey form distributed online through Amazon.
- Data was posted to Kaggle by user Mobius: https://www.kaggle.com/datasets/arashnic/fitbit

Dataset 2 – Exercise lab

- Exercise physiology lab data collected from clinics across the USA in 2017, collated by Graeme Malcolm, Data and AI Content Development Manager at Microsoft, Redmond Washington. Data was initially posted as materials for a Microsoft Azure machine learning training course ("Microsoft DAT263x Introduction to Artificial Intelligence (AI)" - https://www.youtube.com/watch?v=21nsGTBFE4M). Data was reposted to Kaggle by user Fernando Fernandez in 2018: https://www.kaggle.com/datasets/fmendes/fmendesdat263xdemos/data?select=exercise.csv
- 15000 individuals from across USA, representing multiple age groups and an even division of male and female participants
- Feature columns:
    - **User_ID** – unique participant identifier
    - **Demographic/ personal information**
        - Gender, Age, Height, Weight
    - **Exercise metrics**
        - Duration – (length of exercise session in minutes)
        - Heart_Rate (bpm)

- Body_Temp – (skin surface temperature in ˚C)

Dataset 1 – Fitbit

1. Data acquisition and cleaning

[Section 1 – jupyter notebook]

** A note on the jupyter notebook - The python code is presented as one jupyter notebook, which is quite long. However, it has been partitioned into sections and subsections to be easily navigable if opened in PyCharm. Go to the 'Structure' section to be able to click through the headings/ subheadings.*

- Python was used for initial data processing/ transformation, and for regression models.
- Data was in the form of 18 .csv files.
- **Primary files of interest**
  - related to activity and sleep metrics aggregated by day. Hourly steps were also considered in the analysis.
  - dailyActivity_merged.csv → activity_daily data frame
    - Aggregated by day for each participant Id: TotalSteps, distance travelled by ActivityLevel, minutes spent by ActivityLevel, daily Calories burned
  - sleepday_merged.csv → sleepday data frame
    - Aggregated by day for each participant Id: TotalMinutesAsleep, TotalTimeInBed, TotalSleepRecords
  - hourlySteps_merged → steps_hourly data frame
    - Aggregated by hour for each participant Id: StepTotal
  - .csv data relating to heart rate and weight_log was also imported, but was not utilised in the analysis. This will be covered in the general summary.

Participants by metric

- **Activity, steps, calories burned: 33 participants**
  - n>30, which is a reasonable foundation for data analysis
    - we will be able to generalise our results to the larger population in accordance with CLT
- **Sleep: 24 participants**
  - n<30 Smaller sample size than we would like
  - but sleep is a very important metric for this analysis therefore we will include this data
- **Heart Rate: 14 participants**
  - n<30 Significantly smaller sample size
  - I would very much like to include HR data in this analysis (particularly the impact of higher activity levels on resting HR
  - In this case, we may need to favour Dataset 2 – Exercise Lab to gain insights from HR metric that can be generalised outside the sample.
- **Weight: 8 participants**
  - Sample size is too low to include.

- This is disappointing as I would have liked to include BMI information in the analysis as a general health metric. Again, we will need to favour Dataset 2 to include weight/ height information in our analysis.

Data cleaning summary

- activity_daily data frame
  - key for this data frame, uniquely identifying each row:
    1 row = 1 ActivityDate for 1 individual participant Id
  - renamed column: ModeratelyActiveDistance --> FairlyActiveDistance (for consistency)
- 3 duplicate rows removed from sleepday data frame
- weight_log has 65/67 nulls in the 'Fat' column --> discard this column
  - only 8 participants recorded their weight, so we will most likely not use this data frame
- all other data frames contained no nulls/ dupes
- all of the date/ time columns were converted from strings to datetimes
- Joined activity_daily and sleepday data frames on 'Id'→'Id' and 'ActivityDate'→'Sleepday. Produced one combined activity_sleep_daily data frame.

2. EDA

[Section 2 – jupyter notebook]

Table 1: Summary Stats activity_daily data frame

SummStat	TotalSteps	TotalDistance	VeryActiveDistance	FairlyActiveDistance	LightActiveDistance	SedentaryActiveDistance	VeryActiveMinutes	FairlyActiveMinutes	LightlyActiveMinutes	SedentaryMinutes	Calories
count	940.00	940.00	940.00	940.00	940.00	940.00	940.00	940.00	940.00	940.00	940.00
mean	7637.91	5.49	1.50	0.57	3.34	0.00	21.16	13.56	192.81	991.21	2303.61
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
max	36019.00	28.03	21.92	6.48	10.71	0.11	210.00	143.00	518.00	1440.00	4900.00
std	5087.15	3.92	2.66	0.88	2.04	0.01	32.84	19.99	109.17	301.27	718.17

- Mean daily calorie burn for an individual = 2303.61 cal/day
- Most exercise time was performed at the Lightly Active level (LightlyActiveMinutes = 192.81 : around 3.2 hours of light activity per day). Mean VeryActiveMinutes was comparatively low (21.16 min)

Table 2: Summary Stats for sleep metrics

SummStat	TotalSleepRecords	TotalMinutesAsleep	TotalTimeInBed
mean	1.12	419.17	458.48
min	1	58	61
max	3	796	961
std	0.35	118.64	127.46

- Mean sleep time: 419.17 mins = 6.986 Hrs. This is slightly below the 7 hours sleep per night recommended for good health.
- Min sleep time was under 1 hour, which is concerning, but could be due to an incomplete record.

Relationship of TotalSteps → Calories burned (across all participants)

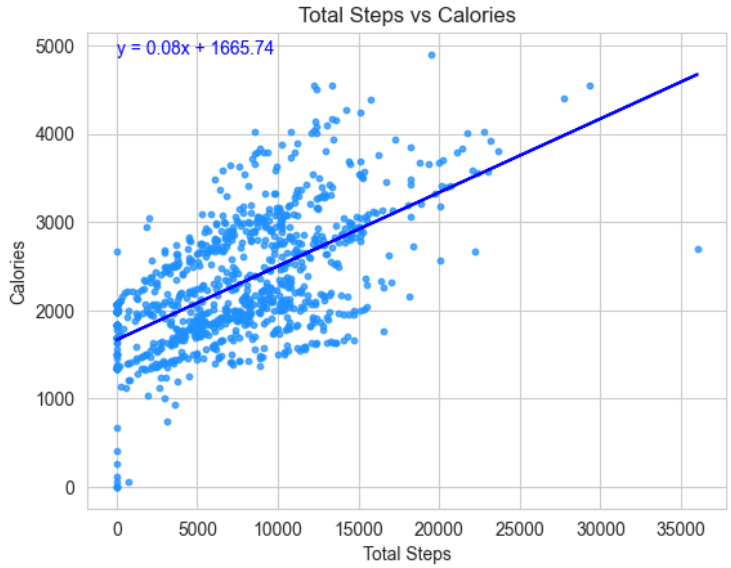


Figure 1: Scatterplot of TotalSteps vs Daily Calories burned across all participants.

From the scatterplot, there appears to be a moderate positive correlation between TotalSteps and Calories. This confirms what we would expect: Higher activity levels (steps) correlate to more calories burned per day. Calculating this correlation: r = 0.592.

For this fit line: $r^2 = 0.35$, indicating that a model with TotalSteps as its only predictor will account for approximately 35% of the variation in Calories.

Note the calories burned per day with no steps taken – Basal Metabolic Rate (BMR) is the number of calories burned to maintain basic functions. Hence, the y-intercept of the fit line for TotalSteps vs Calories can be taken as an estimate of BMR. Average BMR estimate across all participants = 1665.74 cal/day. This is in line with expectations: BMR typically ranges between 1000 - 2000 cal/day for an average adult, with the mean differing according to gender (women ~ 1400 cal/day, men ~ 1800 cal/day).

Relationship between minutes spent at different ActivityLevels and Daily Calories Burned

Fitbit divides users' time into minutes spent at 4 different activity levels (VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, and SedentaryMinutes). The plot below explores the

relationship between accumulated minutes at a certain ActivityLevel and daily Calories burned.

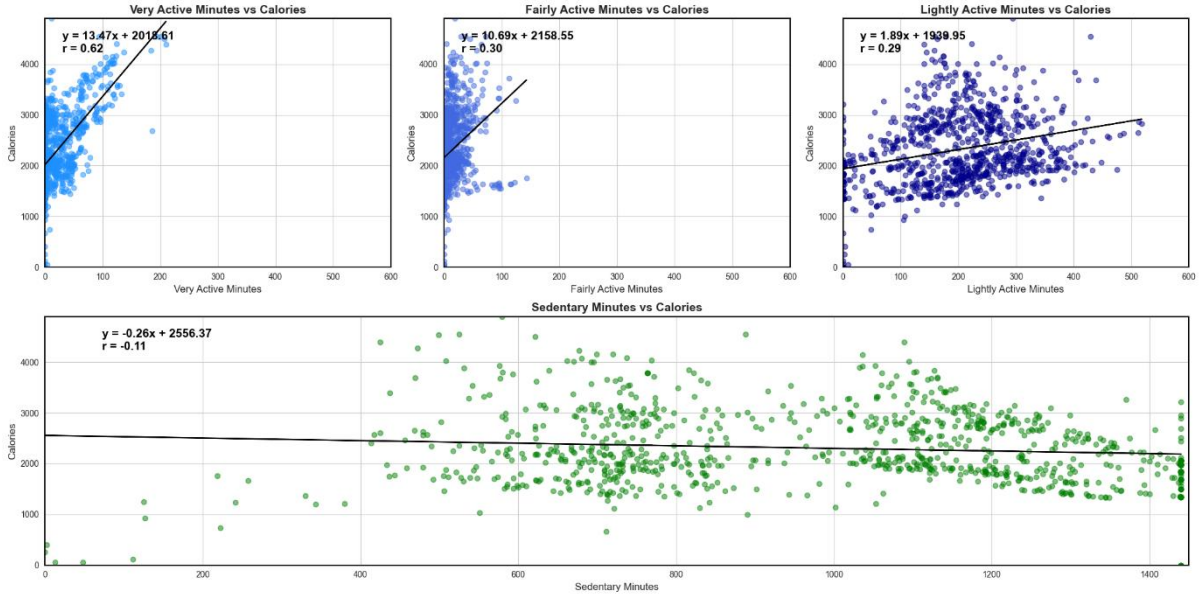


Figure 2: ActiveMinutes vs Calories at 4 different intensity levels.

Note: the sedentary minutes graph is on a different scale to the others because higher values of sedentary minutes were recorded for all participants compared to all other activity intensity levels.

As we would expect higher levels of VeryActiveMinutes correlated to a greater increase in calories burned than the other intensity levels. The slope of the positive relationship to calories was progressively less as the intensity level lowered. More time spent at lower intensity levels correlated to lower calorie burn rates. Less intensity = slower calorie burn
As the intensity level lowered, the relationship to calorie burn was also less tight → the r - value decreasing as intensity level decreased

Note that the fit line for Sedentary Minutes vs Calories was almost flat, confirming that more time spent inactive did not burn more calories.

Basal Metabolic Rate:
Note that once again none of these graphs go through the origin due to Basal Metabolic Rate (calories burned while at rest).

Active minutes vs active distance

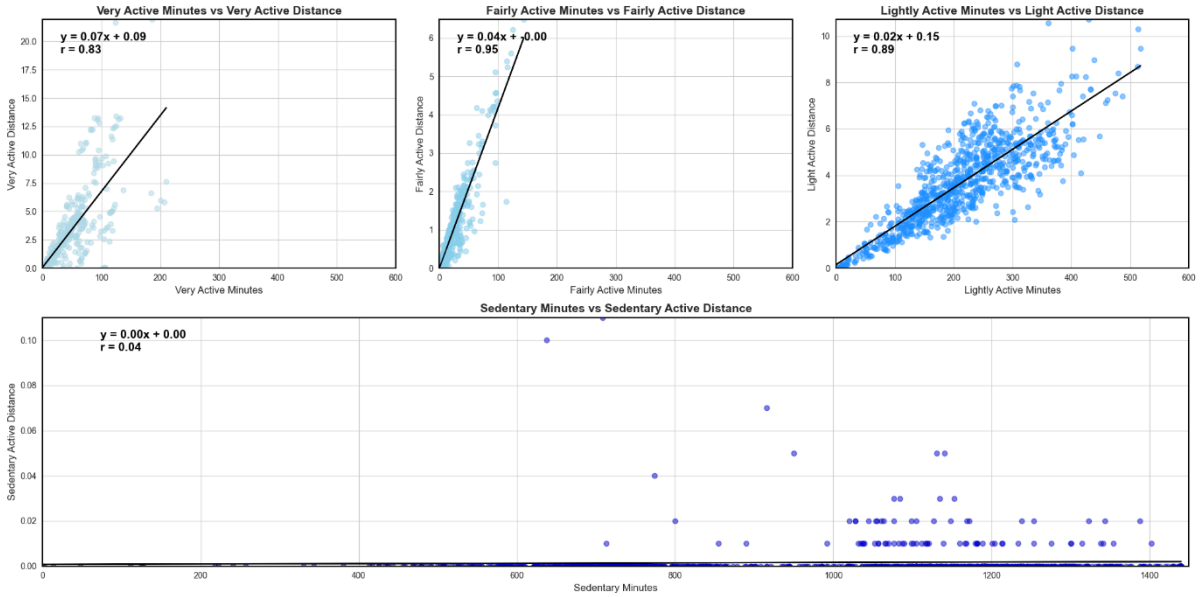


Figure 3: ActiveMinutes at 4 intensity levels vs the corresponding ActiveDistance metric.

Fitbit also classifies distance travelled into the same categories used for ActiveMinutes. The above graph displays the distance travelled in a day for each of the activity levels (Very/ Fairly/ Lightly/ Sedentary) plotted against the number of minutes spent at that activity level for the day.

The results are what we would expect, more time active = more distance travelled. SedentaryMinutes vs SedentaryActiveDistance is basically flat. No distance is travelled when sedentary.

Note that unlike the ActiveMinutes vs Calories graphs, these graphs do go through the origin: 0 active minutes --> 0 distance travelled.

Interestingly, for the same number of minutes, participants travelled more distance with FairlyActiveMinutes vs VeryActiveMinutes (but still burned less calories). The relationships between ActiveMinutes and ActiveDistance are much tighter (compared to ActiveMinutes vs Calories). r - values (0.83, 0.95, 0.89). Strong positive relationships.

How did participants divide their time between the 4 activity levels?

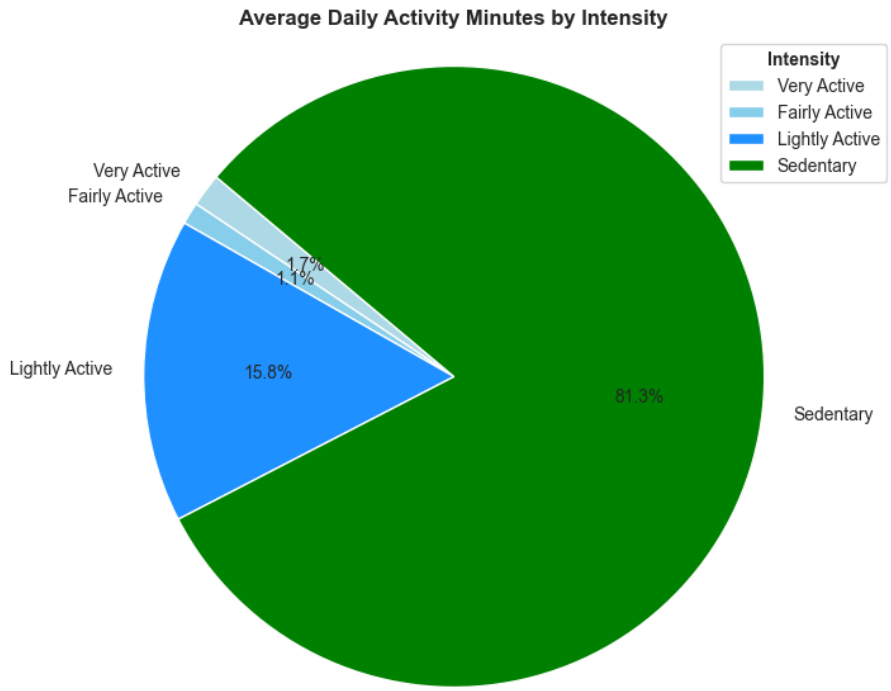


Figure 4: Breakdown of minutes of the day by Activity Level, averaging across all participants.

Participants overwhelmingly spent most of their time sedentary. The majority of exercise was light activity. FairlyActiveMinutes and VeryActiveMinutes only made up a small fraction of their day (1.1% and 1.7% respectively).

Distribution of Average Steps Per Day across all Participants

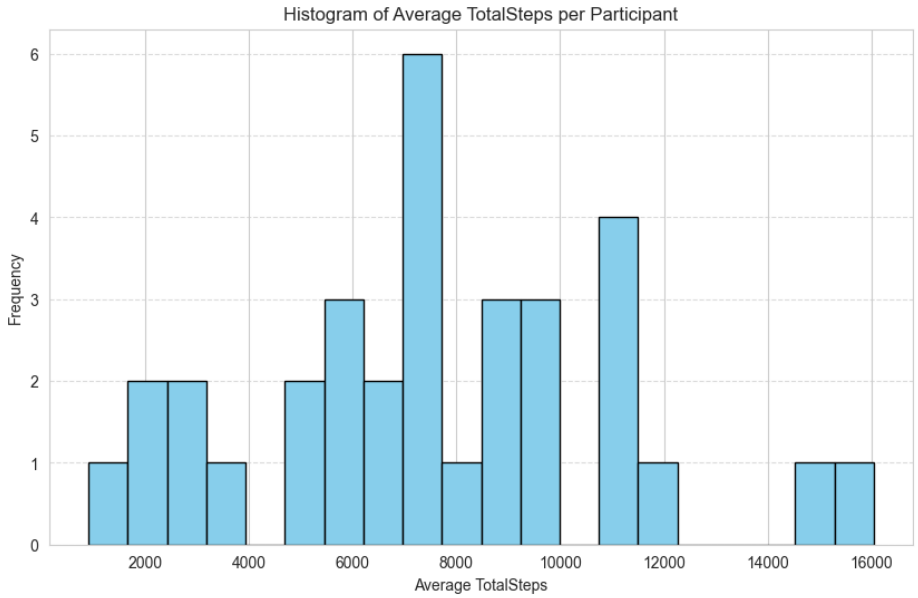


Figure 5: Distribution of Average TotalSteps Per Day. Average daily steps by participants formed a relatively normal distribution, with the highest frequency of average step count in the 7500 - 8000 range. This is lower than the recommended 10,000 step per day target recommended for fitness benefits.

Correlation matrix – Activity and Sleep Metrics

Creating a correlation matrix confirms an association between certain features, and will help us determine if we can include these features in a predictive model.

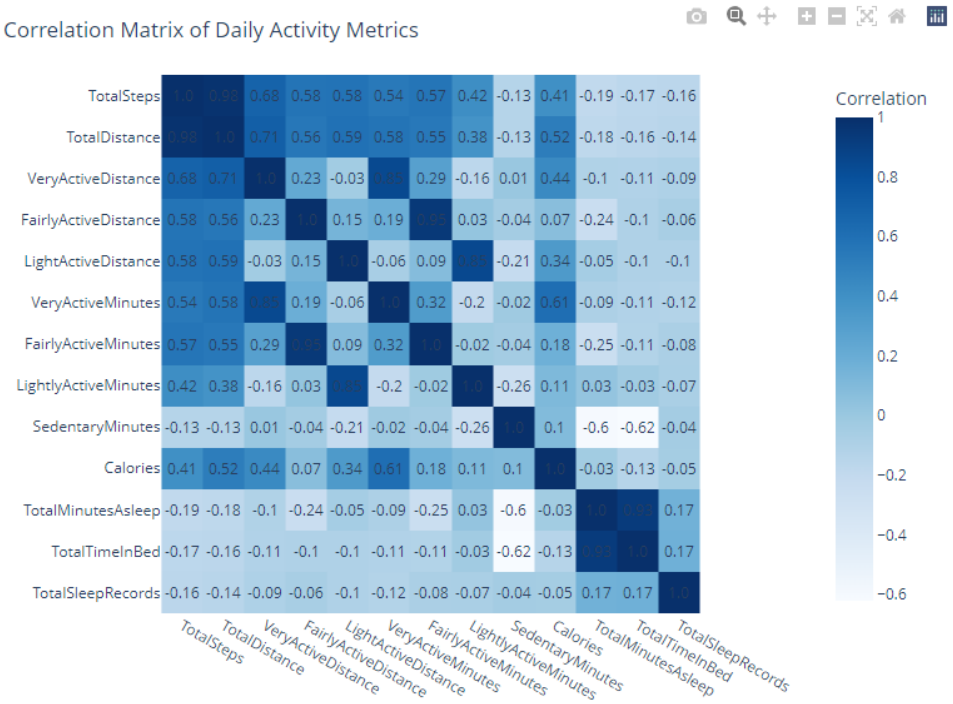


Figure 6: Correlation matrix – activity and sleep features.

Activity metrics:

Focusing on Calories, since we are looking for features that could be potential predictors for this variable:

- Moderate positive correlation TotalSteps --> Calories (r = 0.41)
- Moderate positive correlation VeryActiveMinutes --> Calories (r = 0.61)
- Moderate positive correlation TotalDistance --> Calories (r = 0.52)
- Moderate positive correlation VeryActiveMinutes --> TotalSteps (r = 0.61)
- Weak negative correlation between SedentaryMinutes and [TotalSteps, LightActiveDistance, LightlyActiveMinutes]

As we would expect ActiveMinutes and TotalDistance correlate positively with calories. However, we know the Fitbit algorithm uses ActiveDistance metrics to calculate Calories - i.e. these features are already mathematically related. Due to this auto-correlation these distance metrics are not ideal to include as predictors. A model using these will not tell us anything new. The intensity levels of ActiveMinutes are categorised using ActiveDistance, so we run into a similar problem with these metrics.

Sleep metrics:

- TotalMinutesAsleep and TotalTimeInBed both had a strong negative correlation with SedentaryMinutes
- This indicates that more sleep/ more time in bed led to less sedentary behaviour during the day.

Pairs plot using R: Activity and Sleep

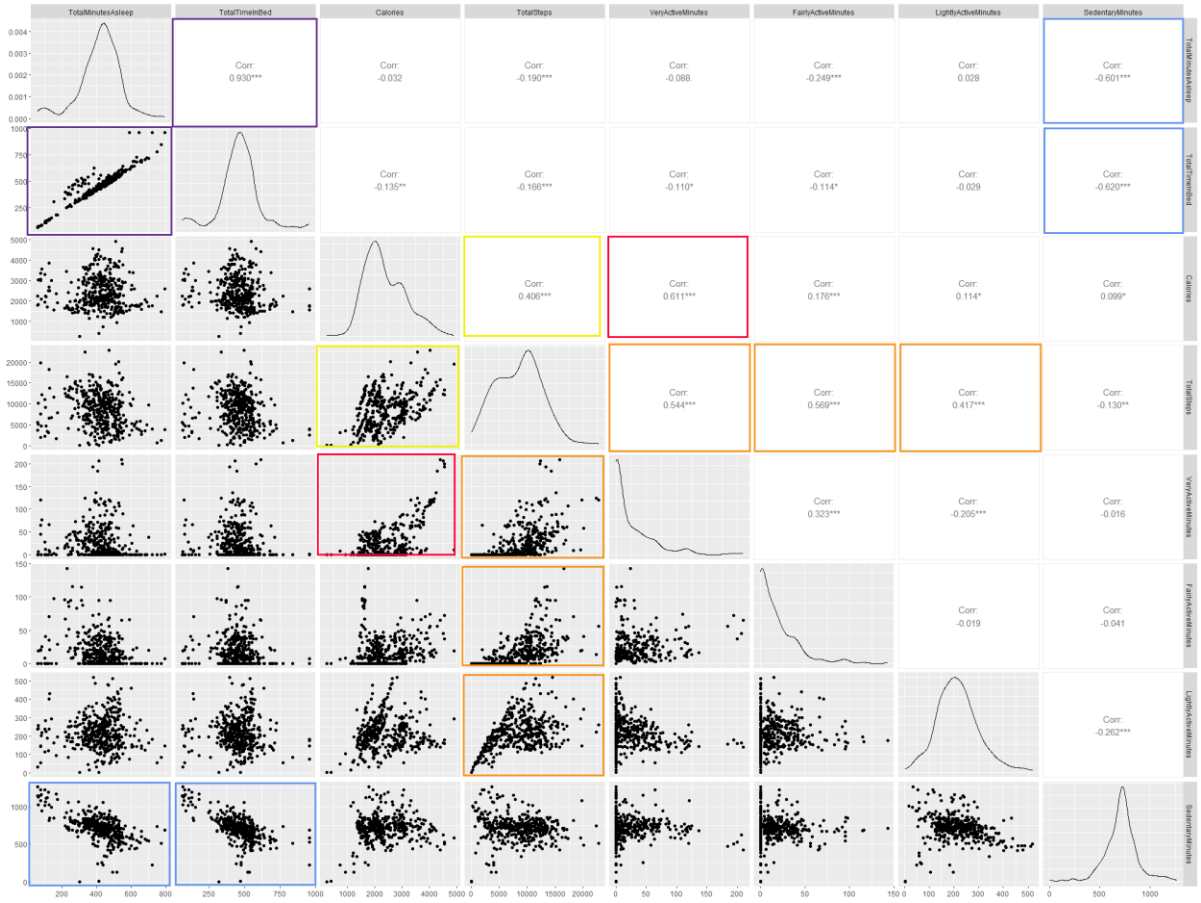


Figure 7: Pairs plot, activity and sleep metrics, showing scatter matrix of the various features and correlations.

Correlations/ Densities:

[Purple] The strongest correlation was between TotalTimeInBed and TotalMinutesAsleep - this indicates that time in bed converted effectively into sleep time.

[Density plot - TotalMinutesAsleep] This was a relatively normal distribution although there is a slight increase in density at the extreme lower end - there were some participants who recorded extremely low total sleep times. The mean TotalMinutesAsleep was 419 min = 6.98 hours. This is slightly under the recommended 7 hours sleep recommended for adults.

[Blue] Strong negative correlation between both TotalMinutesAsleep, TotalTimeInBed and Sedentary minutes. More sleep --> less sedentary during the day.

[Red] Moderate-Strong positive correlation between VeryActiveMinutes and Calorie burn. The most intense activity had the greatest contribution to calorie burn. However, looking at the density plot for VeryActiveMinutes confirms that this level of intensity occurred very infrequently for most participants.

[Density plots - VeryActiveMinutes, FairlyActiveMinutes, LightlyActive minutes]
Only LightlyActiveMinutes appeared closer to a normal distribution --> with mean 193 mins situated in the center of the data. The other higher intensity levels were very right skewed, indicating that participants rarely exercised at this intensity.

[Orange] Moderate correlation between Very/Fairly/LightlyActiveMinutes and TotalSteps. Interestingly, even though the r-value is lower (0.417), the relationship between LightlyActiveMinutes and TotalSteps looks "tighter" than the others - i.e. TotalSteps seems to increase more consistently with LightlyActiveMinutes. Would this be because for most participants, lighter exercise was the predominant activity type (besides sedentary)?

[Yellow] Moderate positive correlation (r = 0.406) between total steps and calories. Would this indicate that intensity of exercise (VeryActiveMinutes having a higher r-value 0.611) has an even greater contribution to calorie burn than simply net amount of exercise?

[Density plot - SedentaryMinutes] Relatively normal distribution. But note that the spread is much narrower for SedentaryMinutes, indicating that the amount of time spent sedentary amongst the participants was bunched up around similar values close to the center. Mean SedentaryMinutes = 991 min

---

[Density plot - Calories] Calories burned was slightly right skewed with a bimodal hump close to the center of the distribution. i.e. there was a smaller secondary peak of participant who were burning slightly more calories than the mean. Mean daily calorie burn = 2304 Cal/day

How frequently were participants logging their daily activity?

- To what extent were Fitbit users actually wearing their device during the day?
- This will be calculated by summing activity minutes:
    - activity_mins_sum = VeryActiveMinutes + FairlyActiveMinutes + LightlyActiveMinutes + SedentaryMinutes
    - There are 1440 minutes in 1 day
    - Therefore percent_activity_logged = activity_mins_sum/1440*100

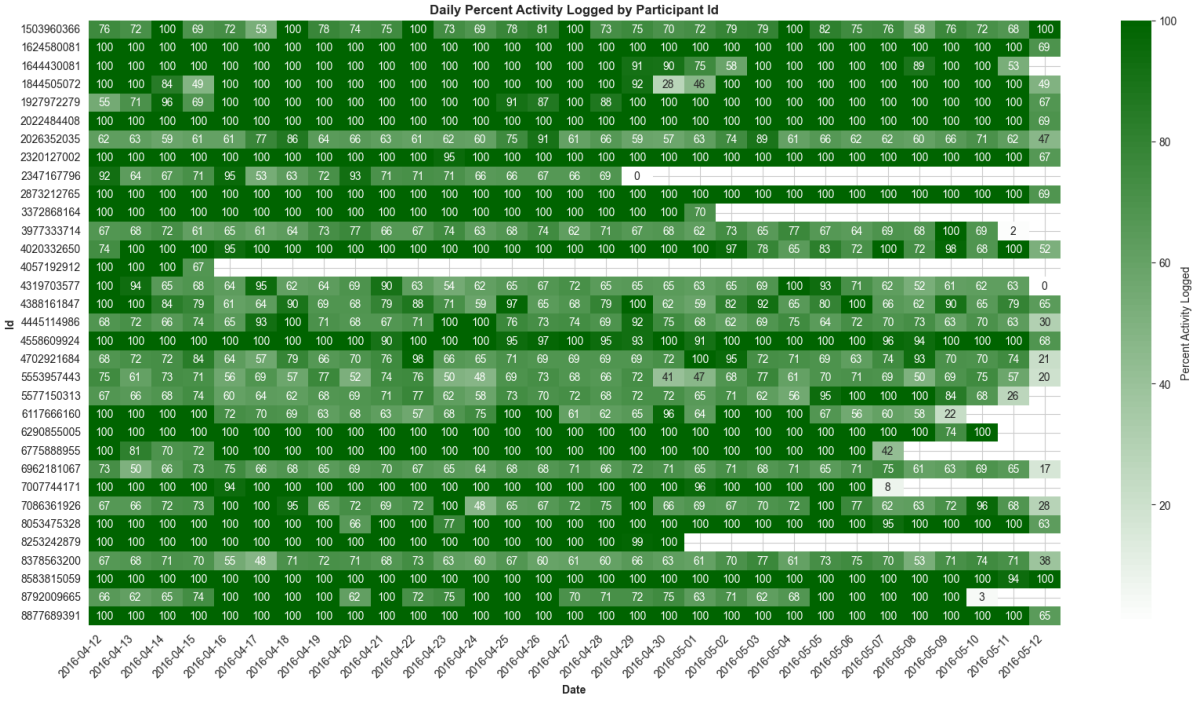


Figure 8: Percentage of minutes logged for each participant Id, for each of the 31 days of the study.

The above figure shows the % of minutes in the day that was logged as either (VeryActiveMinutes/ FairlyActiveMinutes/ LightlyActiveMinutes/ SedentaryMinutes). It can be seen from this that a significant proportion of participants logged nearly all of their days using their wearable device. Overall, there was a predominance of logged days. Participants logging partial days generally did so for most days in the study. There were no participants that logged all of their days 100%.

There is an interesting pattern here where those logging partial days would do so in blocks – logging a series of partial days, then having a full day or two, before going back to partial days.

Those that stopped logging completely never went back to it. They presumably dropped out of the study.

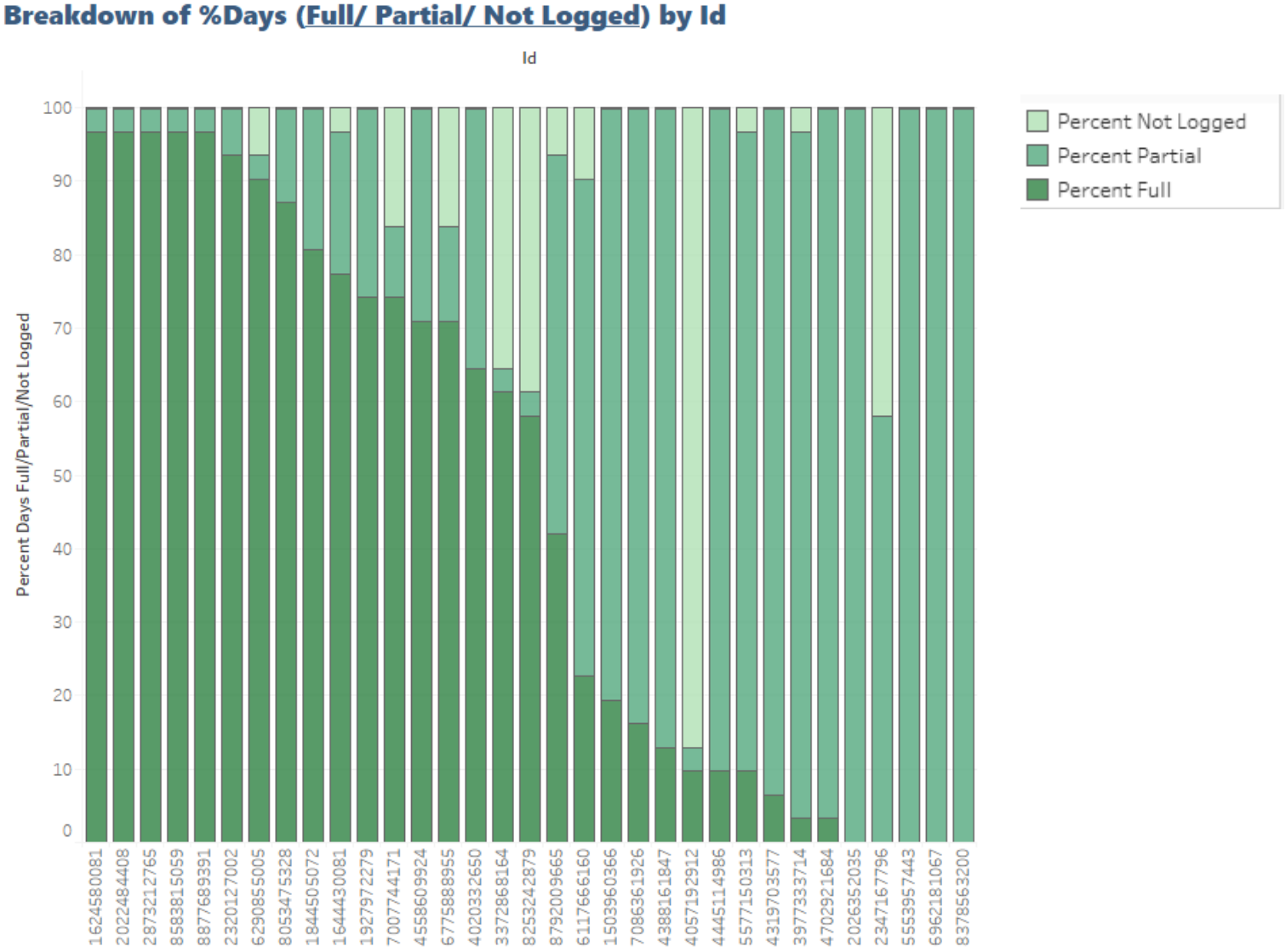


Figure 9: Percent Days (Full/ Partial/ Not Logged) by Id. Interactive version available in Tableau workbook.

Above shows the breakdown of Full/ Partial/ Not Logged days by participant Id. It appears that there is a roughly even split between those logging a higher proportion of full days and those logging more partial/ none.

As the percent_minutes_logged metric seems to separate the participants along a continuum, this may be a good candidate for a predictor in a regression, if it correlates with Calories.

How did activity vary across days of the week?

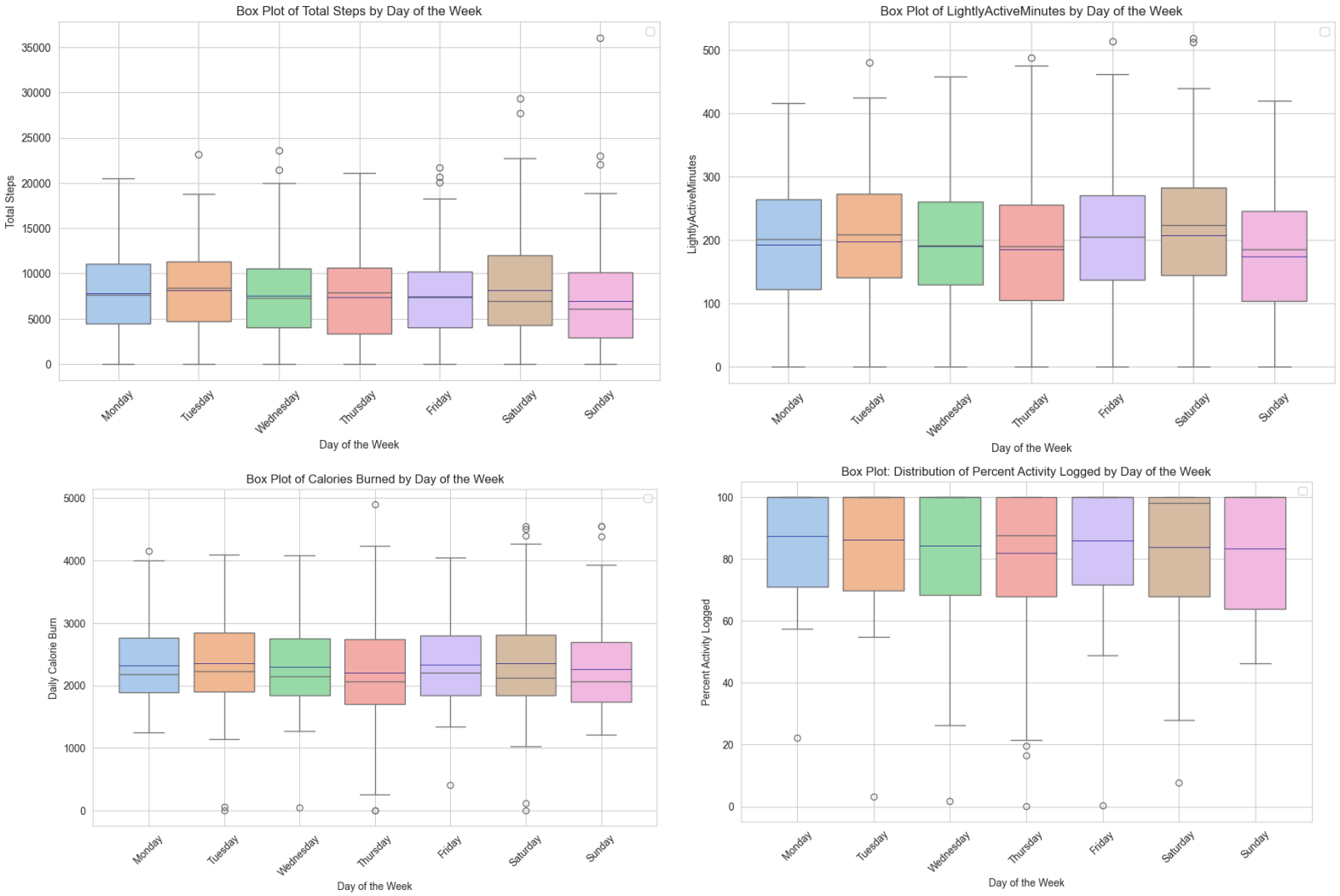


Figure 10: Distributions – (Total Steps, LightlyActiveMinutes, Calories Burned, Percent Activity Logged) by day of week. Mean of distribution in blue.

- Steps by day of week
  - No significant pattern of change across the week – IQR (middle 50% of data) overlaps for all days. Based on this visual, we cannot say that TotalSteps, grouped by days of the week are significantly different.
  - Saturday – highest mean steps by day of week and greater spread of data.
  - All of the distributions are right-skewed, indicating more observations clustered at lower values.
- LightlyActiveMinutes by day of week
  - Follows similar pattern to steps
- Calories burned very similar across days.
- Percent activity logged distributions
  - All left skewed.
  - Widest spread of data on Wed and Thurs – indicating that some outlier individuals had particularly low logging activity mid-week
  - Otherwise IQRs overlap, mean logging is not significant across days of week.

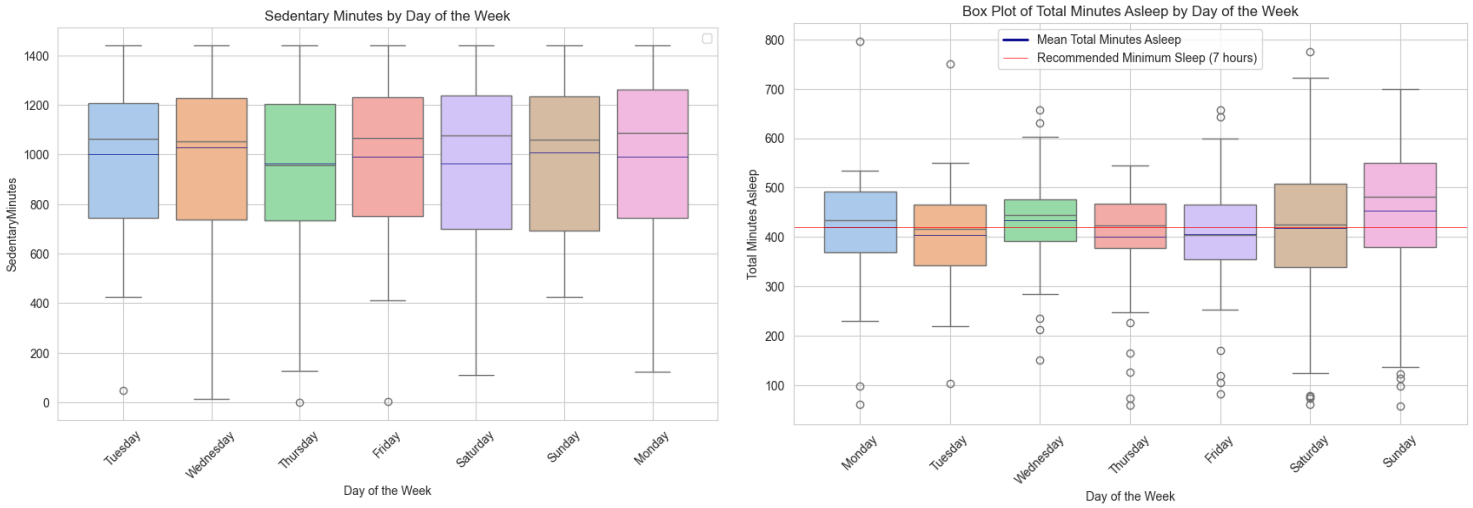


Figure 11: Distributions – SedentaryMinutes, TotalMinutesAsleep by day of week.

- Sedentary mins by day of week
  - All distributions left-skewed
  - Otherwise no discernible weekly pattern from the visual– means are similar across days of week
- TotalMinsAsleep by day of week
  - The mean total sleep minutes per day was below 7 hours on Tue, Thurs, Fri. Sleep was generally better on the weekends, though there was a greater spread of values. For Sundays, spread of total sleep minutes was left skewed with mean > recommended 7 hours.

Weekday vs weekend

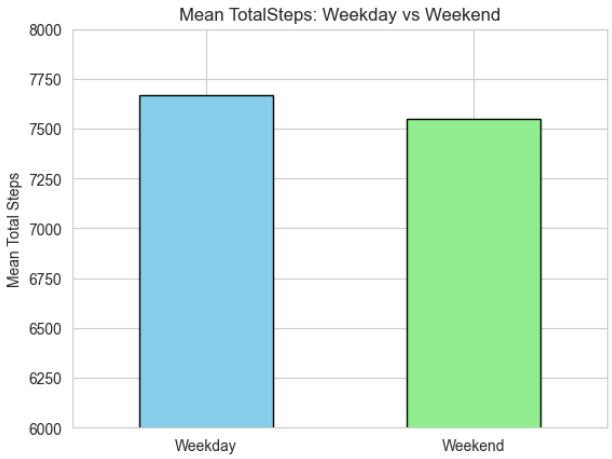


Figure 12: Mean total steps weekday vs weekend across all participants. Participants were generally more active during the week, with higher total steps on weekdays.

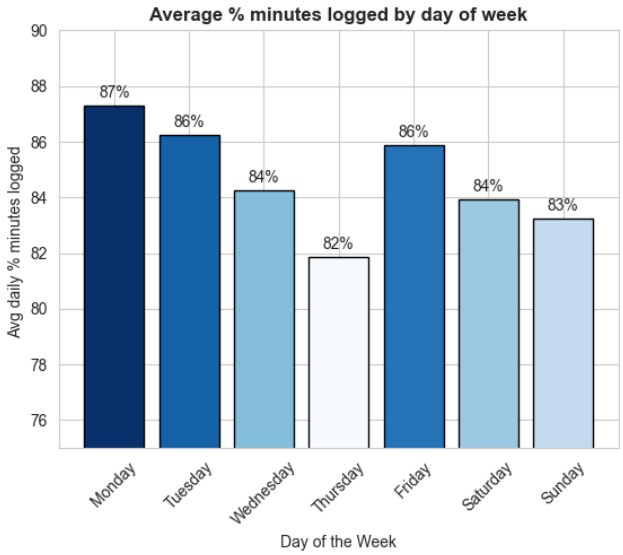


Figure 13: Average percent minutes logged by day of week. Scale of the graph separates the values more dramatically; actually, there is only a small variation in these values (range 82 – 87%). Potentially, this is not a significant difference, but there is a steady decrease in logging from Mon – Thu, increasing on Fri, then decreasing towards Sun.

Activity across hours of the day

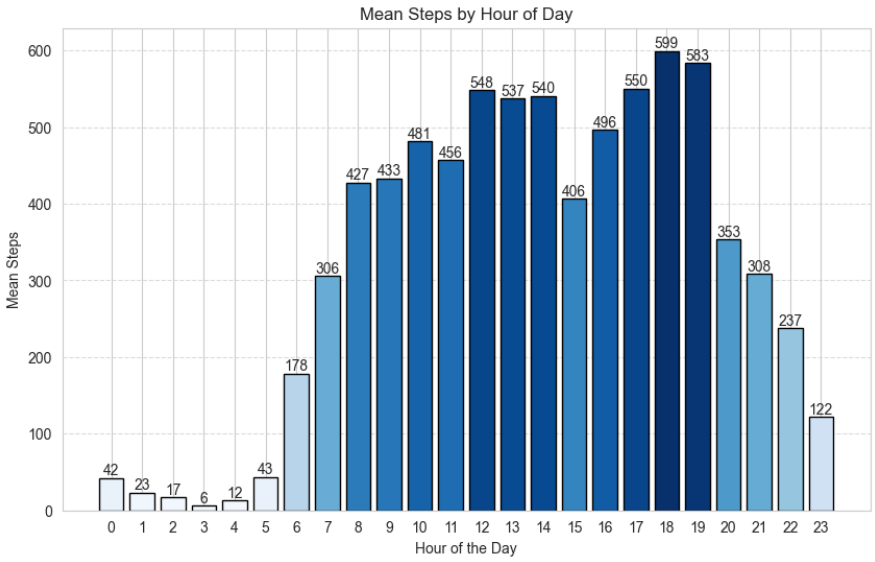


Figure 14: Mean Steps by hour of day. Activity increases from 6am, reaching peaks from 12-2pm (~540 steps) and 5-7pm (550-599 steps), before dropping sharply at 8pm and continuing to decrease. There

was a small amount of activity after midnight that may be attributed to restlessness/ getting out of bed. The lowest activity was at 3am.

3. Preparing explanatory and response variables, and regression analysis

[Section 3A,B,C in jupyter notebook]

- The objective is to build models linking activity metrics to Calories burned/ day and general health metrics.
    - Get mean Calories burned per day by participant Id
    - This will allow us to observe any correlations of Calories with percent_month_minutes_logged, and any other general health metrics we might create (which are also aggregated by Id)
- There was little demographic information included for each participant (weight_log had too many nulls. Age and gender were not included. Before switching to the Exercise Lab dataset, it is worthwhile to see if any general health metrics may be inferred purely from the activity data present.
    - We saw earlier that BMR may be estimated using the intercept of a regression line on a plot of TotalSteps vs Calories.
    - This calculated BMR stat could then be used as a response variable in further regressions.
- Check the correlation between logging frequency and average daily calories burned by participant Id. Logging frequency may be a useful predictor for Calories.
- By the end of this process we should have:
    - Proposed response variables:
        - Average daily Calories burned by Id
        - Average estimated BMR by Id
    - Proposed predictors:
        - Percent_month_minutes_logged by Id
        - Any other activity/ sleep metrics aggregated by Id.

3A. Estimate BMR for each participant by regressing TotalSteps against Calories

- *[process is explained in detail with code in jupyter notebook]*
- Average BMR estimates were obtained along with a 95% confidence interval for each of the 33 participants.
- Statsmodels was used for this regression. p-values for intercepts and coefficients were assessed for significance (<0.05), as was r-squared for each model.
- The focus here is on understanding the relationship between these variables (Calories~TotalSteps). The goal is to later use the BMR estimate as the response

variable in another regression model: we will later attempt to use % minutes logged to predict BMR. So, right now hypothesis testing is enough to assess this current model's performance. For our predictive model later, when we are assessing the model's performance on unseen data, we will need to do a train-test split and calculate RMSE.

- To visualise the results, 95% CI's will be plotted along with the BMR estimate. Percent minutes logged for the month will also be plotted for each participant. This will allow us to see if there is any correlation between logging frequency and BMR

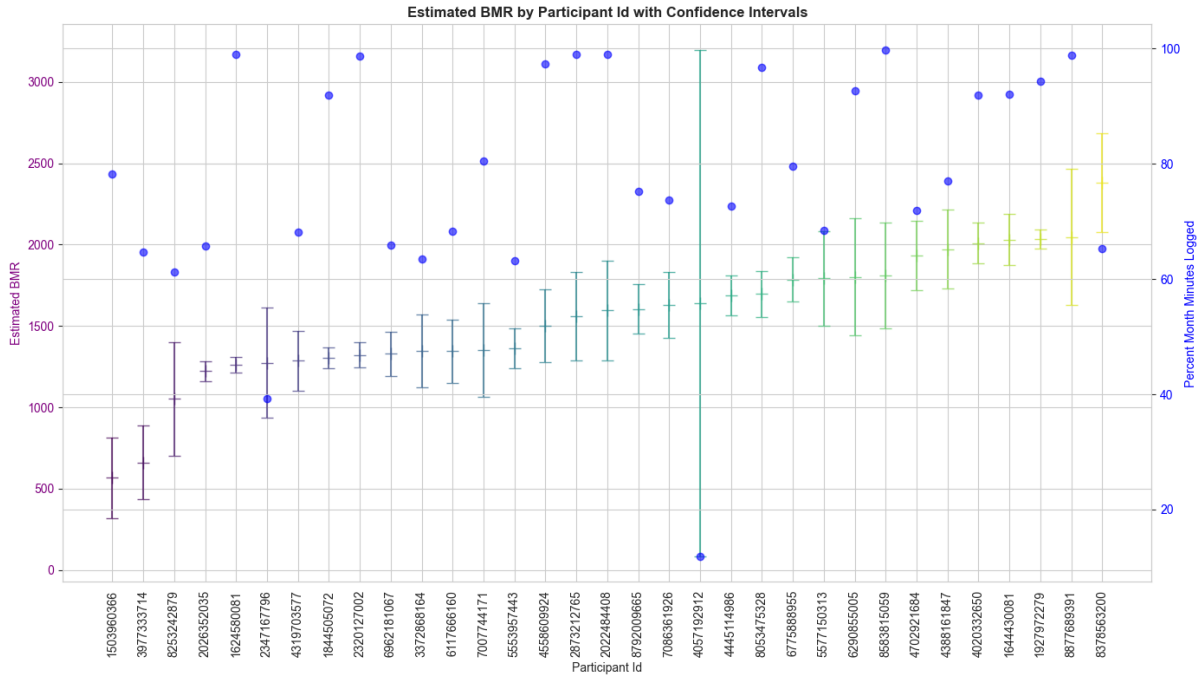


Figure 15. Calculated average BMR values by participant Id, ordered by increasing BMR. 95% CI is represented. Percent minutes logged for the month by Id has also been visualised (blue dots).

- Percent minutes logged appears to be a random scatter with no discernible pattern relating to increasing BMR. The initial hypothesis was that participants who were more diligent with logging would present with better general health metrics, but this appears not to be the case.

- Logging frequency appears not to correlate strongly with BMR. The only exception to this is when logging frequency is so low that it is not possible to accurately estimate BMR using a regression (e.g. participant Id 4057192912).

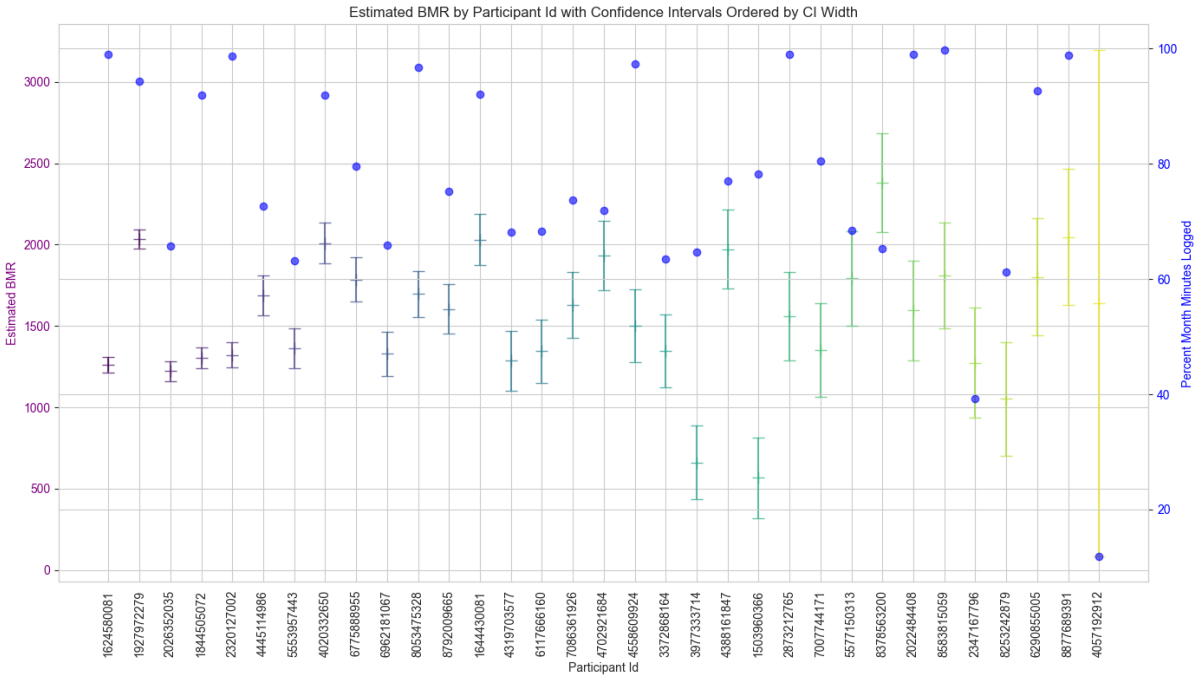


Figure 16: Calculated average BMR values by participant Id, ordered by increasing CI range.

- The above figure was produced to check if there is a relationship between percent minutes logged during the month and the accuracy of the BMI calculation. Again the observations for percent minutes logged presented as a random scatter. Visually, there does not appear to be a relationship between logging frequency and CI width.

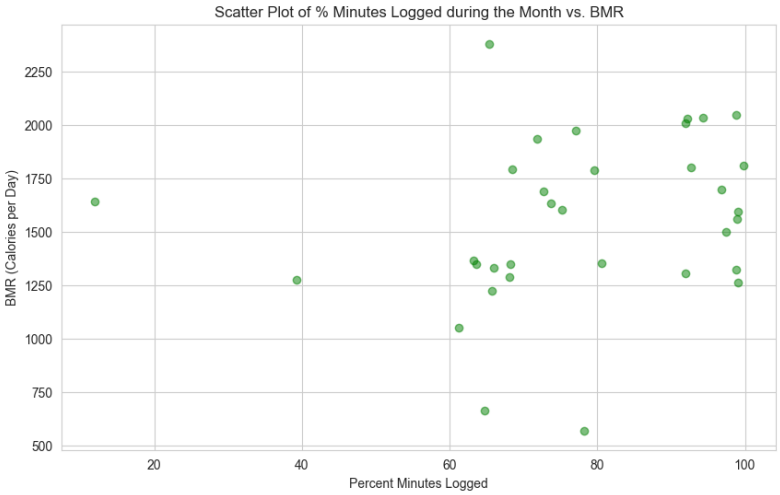


Figure 17: There appears not to be a strong correlation between BMR and percent_month_minutes logged. We will examine this relationship further with a correlation matrix that we produce along with the other features.

3B. Group mean calories burned per day by participant Id

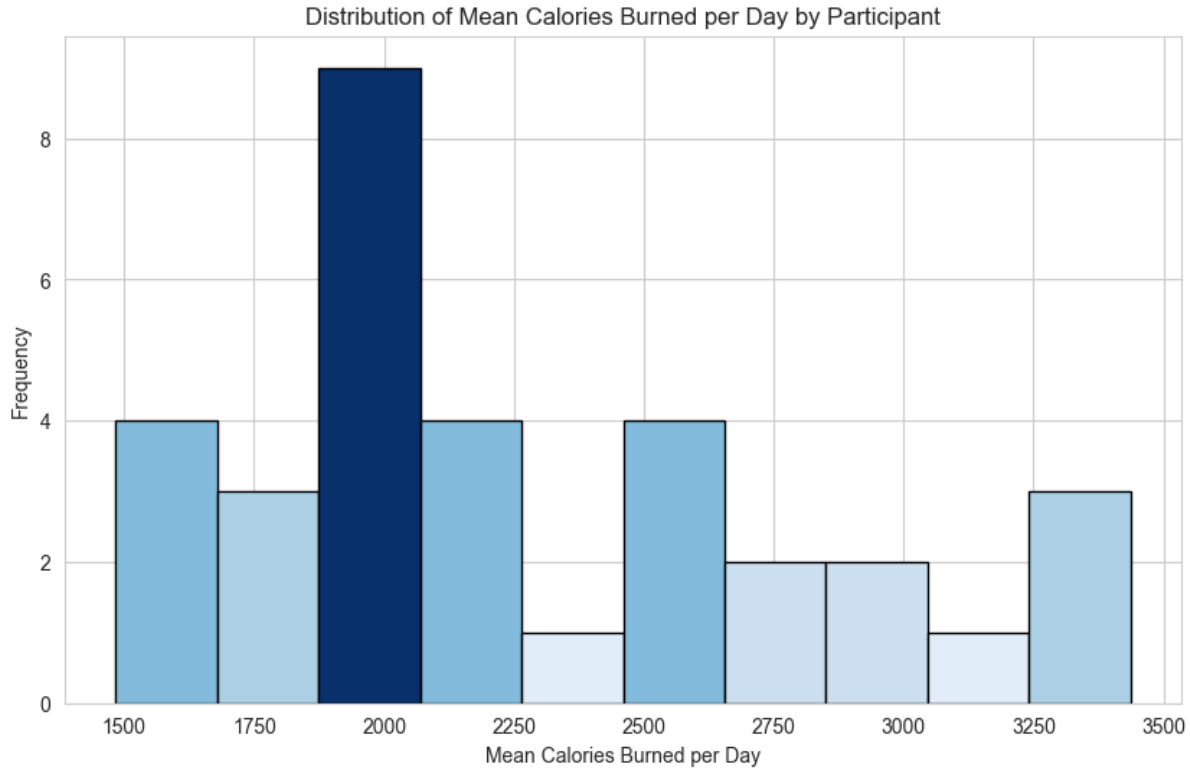


Figure 18: Highest frequency of mean calories burned per day was around 2000 cal/day.

Joined mean calories burned per day by Id with logging data (percent_month_minutes logged) → investigate whether there is a relationship between these two metrics.

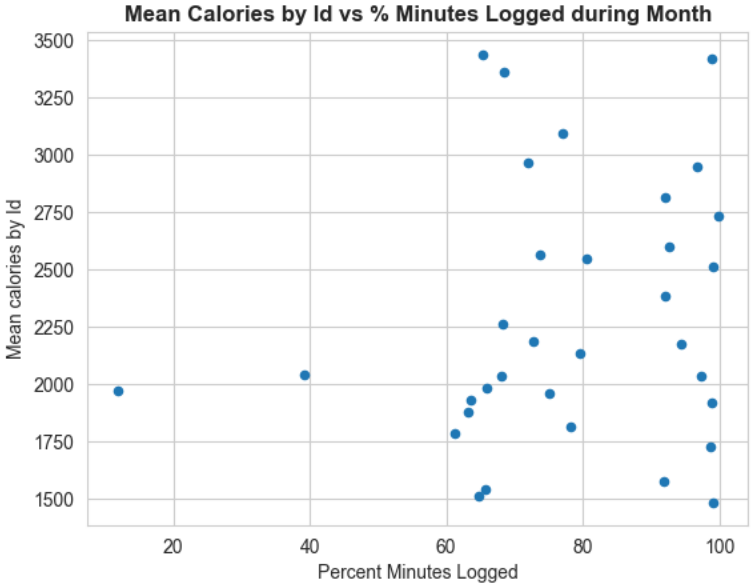


Figure 19: There appears not to be a strong relationship between logging frequency (percent_month_minutes logged) and mean calories burned per day. Note that the majority of % logging fell within the 60 – 80% range for the whole month.

3C. Look for further correlations between metrics

- Since there appears not to be a strong correlation between logging frequency and either of our proposed response variables (Mean Daily Calories Burned by Id, BMR), let's explore some other potential relationships in the data:
    - Look at ratio of active vs sedentary minutes:
        - active_sed_mins_ratio = (Very+Fairly+LightlyActiveMinutes)/SedentaryMinutes
        - Plot active_sed_mins_ratio vs BMR
    - Plot SleepMinutes grouped by Id vs BMR

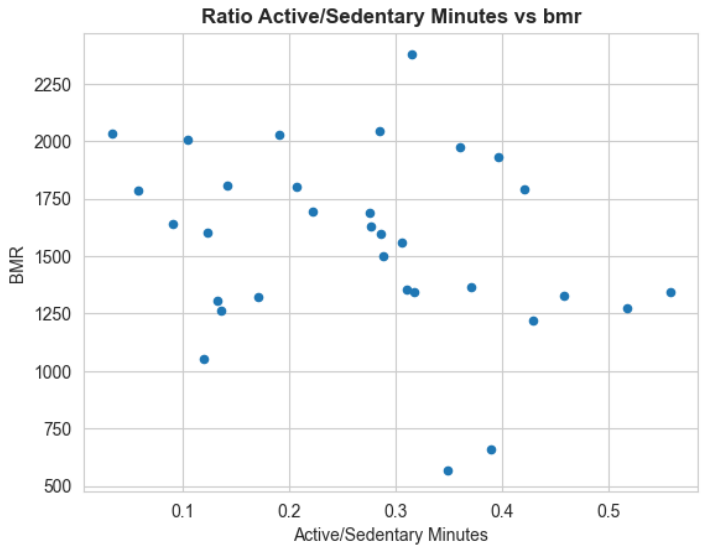


Figure 20: No obvious correlation between active_sed_mins_ratio and BMR .

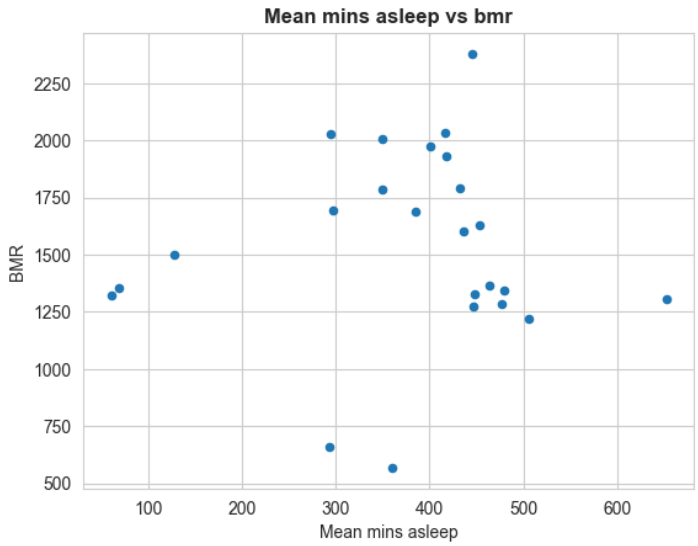


Figure 21: There may be a weak positive correlation between mean mins asleep and BMR.

Correlation matrix with mean metrics grouped by participant Id including:

- Calculated metrics:
  - percent_month_mins_logged (logging frequency)
  - BMR
  - active_sed_mins_ratio (active/ sedentary minutes)
- TotalMinutesAsleep
- Activity metrics, Calories

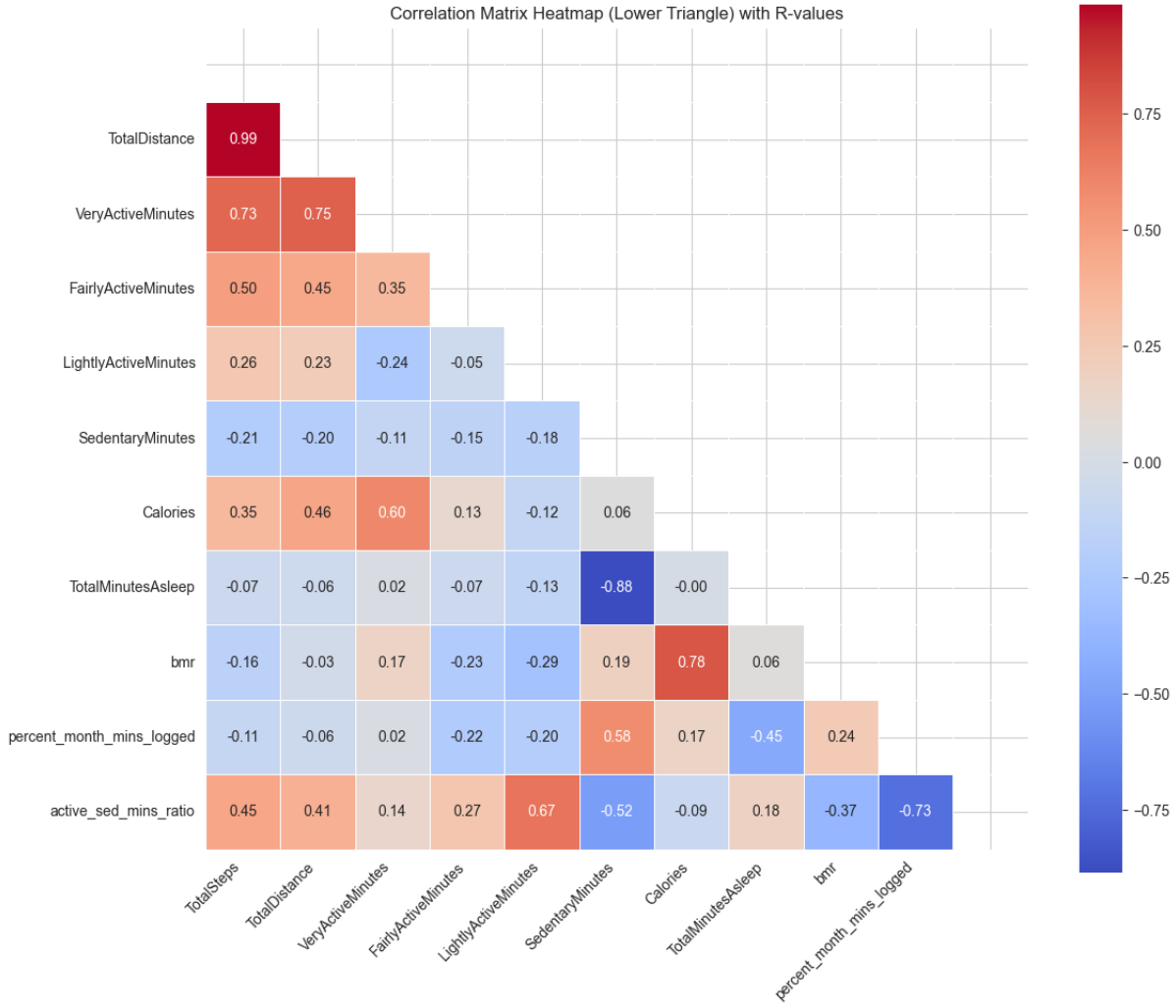


Figure 22: Correlation matrix with activity/ sleep metrics and calculated variables.

- The strong negative correlation between TotalMinutes asleep and Sedentary minutes seen earlier is also represented here.
- The above matrix reveals a problem with our explanatory variables and proposed response variable for the regression:
  - percent_month_mins_logged does not have a strong positive correlation with any of the other metrics except sedentary minutes (r=0.58).
  - This comes as a surprise give our initial hypothesis that more diligent logging would correlate with better exercise success/ health metrics. The EDA implies that we should reject this hypothesis.

    - Reinforcing the positive correlation with SedentaryMinutes, percent_month_minutes_logged has a strong negative correlation with active_sed_mins_ratio (r=-0.73). Those that logged more were just logging more sedentary minutes.
    - There is a weak positive correlation between percent_month_minutes logged and BMR (0.24) – but this is not sufficient justification to use logging frequency as a primary predictor of BMR.
    - In fact, the only strong correlations with our response variables (Calories, BMR) are with features that auto-correlate with these variables
        - i.e. features that already have an established mathematical relationship with Calories or BMR
            - TotalSteps, TotalDistance, VeryActiveMinutes → Calories
            (Calories is derived using the Fitbit algorithm using TotalSteps, TotalDistance, ActiveDistance measures)
            - Calories → BMR
            (BMR was derived from a regression of TotalSteps against Calories earlier in this analysis)
- This presents a problem where the metrics above are less viable candidates for regression analysis. In this case, we will switch over to the Exercise Lab data which contains more demographic information, and also heart rate, skin temperature measurements for an individual exercise session. The dataset also contains more observations (15,000) than our initial Fitbit dataset.
- Note that all the metrics in the Exercise Lab dataset could still be collected by the Fitbit app and wearable, even though they were unavailable in the Fitbit dataset for this study.

3D. Exercise Lab Dataset

Cleaning, dataset exploration, EDA

Summary

- 2 .csv files (calories.csv, exercise.csv)
- 15,000 unique User_IDs, no duplicate records
- No nulls
- Gender of cohort was an even split between males and females
- Calories and exercise data were joined into a single data frame: exercise_calories_merged containing columns:
    - User_ID, Gender, Age, Weight, Height,
    - Duration (length of exercise session -mins)
    - Heart_Rate (mean heart rate during exercise session - bpm)
    - Body_Temp (mean skin surface temperature during exercise - ˚C)
    - Calories (calories burned during this exercise session)
- For our models we will be predicting Calories
    - Therefore the other features will be our explanatory variables.

Table 3: Summary Stats for Exercise Lab Dataset

summStat	Age	Height(cm)	Weight(kg)	Duration(min)	Heart_Rate(bpm)	Body_Temp(Celsius)	Calories
mean	42.79	174.47	74.97	15.53	95.52	40.03	89.54
std	16.98	14.26	15.04	8.32	9.58	0.78	62.46
min	20	123	36	1	67	37.1	1
max	79	222	132	30	128	41.5	314

- Mean duration of exercise session was around 15.5 min
- Mean HR during exercise was 96 bpm – We might infer from this that the sessions were mostly "light exercise". This is below the "aerobic zone" of training (~105 – 140bpm) and closer to the low-end of the "fat burning zone" (~90-120bpm)

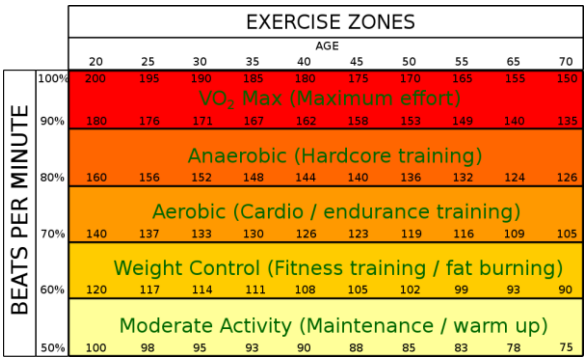
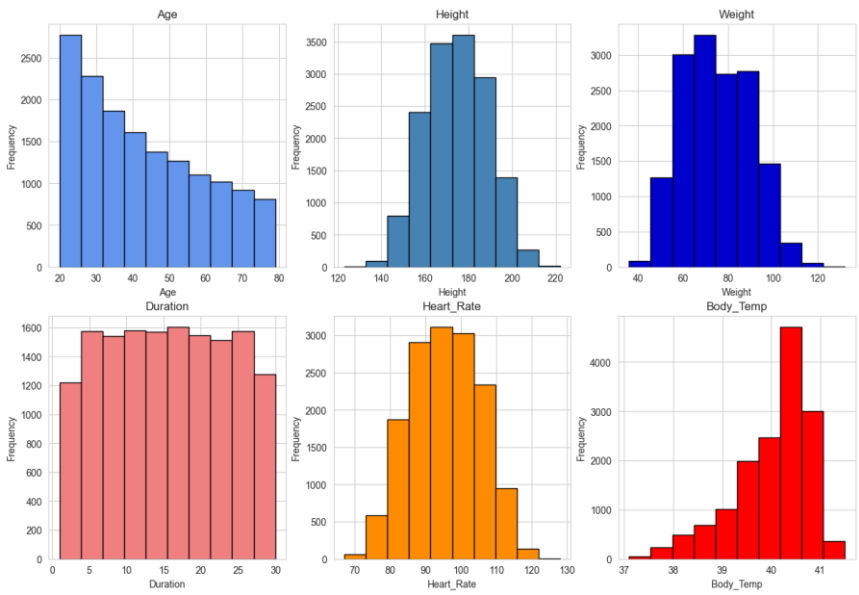


Figure 23 : Exercise Zones
https://fitnesshealth.co/blogs/fitness/14112669-best-heart-rate-to-burn-fat

- Mean body temp during exercise session was 40˚C

- Mean calories burned was 90 cal. From the tables here:
https://www.verywellfit.com/walking-calories-burned-per-minute-3887138
we can see that this is the equivalent of a 160lb (72kg) person walking ~15min at 4mph (6.4km/hr). This is a relatively brisk walking speed.

Distribution of data – Exercise Lab Dataset



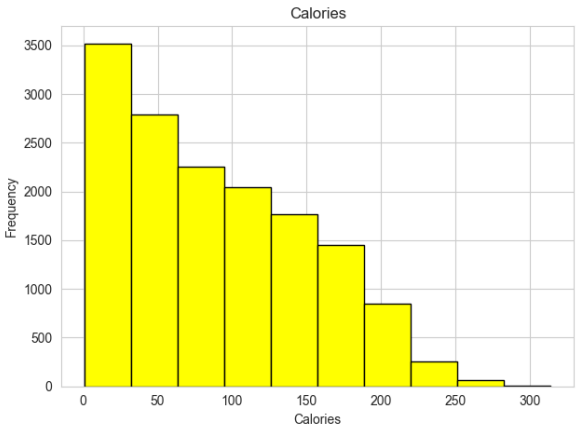


Figure 24: Distribution of data in this dataset followed vey tidy lines.

- There was a smooth drop in count from age 20 →80.
- Calories burned was similarly left-skewed with drop in frequency from 0 → 300.
- Count of observations in duration was relatively evenly spread between 5 – 25min with a drop off at the extreme lower and upper end of the distribution.
- Height, Weight, and average Heart Rate were roughly normally distributed.
- Body Temp was left-skewed with most of the observations clustering around 40-40.5˚C, and a sharp drop-off at the upper end of the distribution.
- This leads me to believe that the dataset has been curated for education purposes – as these results are "too clean".
- Regression analysis on a curated dataset will still provide us with useful information on the relationship between Calories and various predictor variables – we just need to keep in mind that the training data comes from a "controlled environment" when applying this model to out of sample data.

Correlation matrix: Exercise Lab Data

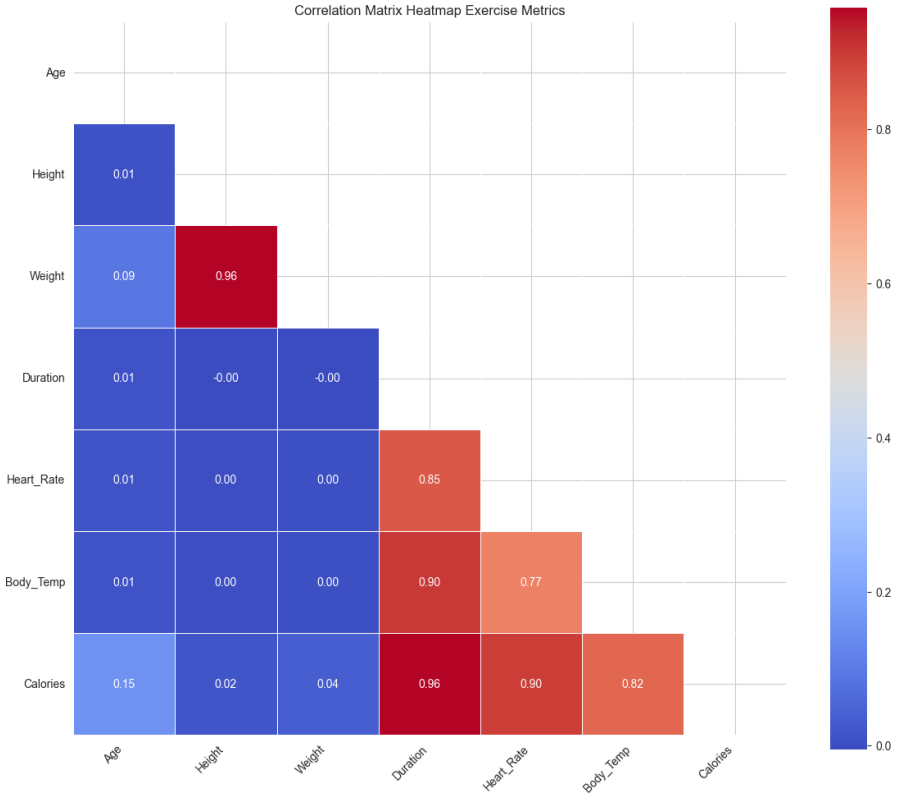


Figure 25: Calories burned during exercise session → strong correlation with:

- Exercise Duration (mins)
- Heart_Rate (bpm)
- Body_Temp (˚C)

These three metrics will be our main predictors. Weight also correlates strongly with height.

Scatter Matrix: Exercise Lab Data

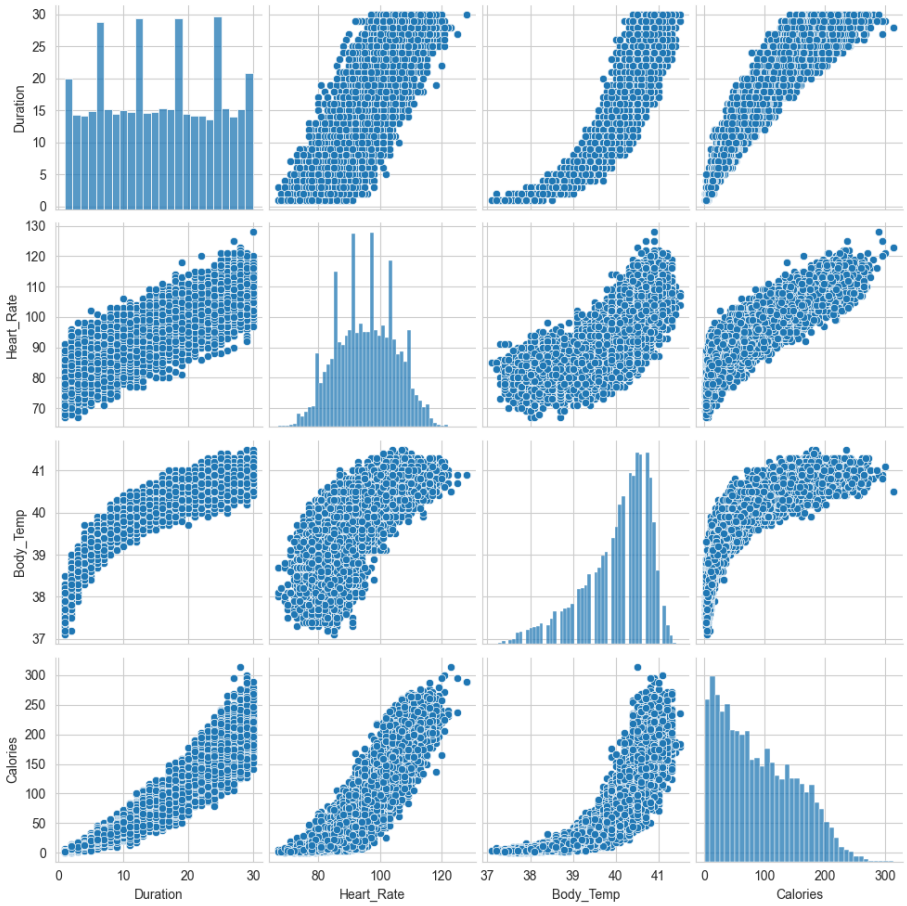


Figure 26: While Duration, Heart_Rate and Body_Temp are positively correlated with Calories, the relationships appear curved rather than linear:

- Duration could have a quadratic relationship with Calories, Heart_Rate (cubic?), Body_Temp - strongly curved (quadratic?)
- We will start with linear models and then progress to more complex models if required
- The scatter matrix appears to confirm the curated nature of the data set -- these relationship look very "clean" with minimal random noise that we would ordinarily see in a raw real-world dataset.

3E. Regression analysis

Dummy variable encoding for Gender category

- Gender column values: male/ female
- One hot encoding was used to convert these values into:
    - Male → true (1)
    - Female → false (0)

(Model 1) Linear model with limited features

Calories ~ Duration, Heart_Rate, Body_Temp

- Predictive model for Calories using only Duration, Heart_Rate, Body_Temp as predictors (features with the highest correlation)
- Model was fitted with sklearn
- Model performance was assessed using train-test split, RMSE calculated and compared with null model RMSE
    - the null model always predicts mean Calories
- Same model was refitted using statsmodels to obtain:
    - p-values for coefficients
    - r^2 value for the model

Results:

model_number	RMSE	r-squared
null	61.86	
1	14.57	0.946

- Model 1 including features ['Duration', 'Heart_Rate', 'Body_Temp']
    - Lower RMSE compared to null model → performs better than null model on out of sample data.

p-values and r-squared:

OLS Regression Results

Dep. Variable:	Calories	R-squared:	0.946
Model:	OLS	Adj. R-squared:	0.946
Method:	Least Squares	F-statistic:	8.801e+04
Date:	Tue, 13 Feb 2024	Prob (F-statistic):	0.00
Time:	01:24:13	Log-Likelihood:	-61374.
No. Observations:	15000	AIC:	1.228e+05
Df Residuals:	14996	BIC:	1.228e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	470.3525	13.806	34.069	0.000	443.291	497.414
Duration	6.6414	0.040	164.632	0.000	6.562	6.720
Heart_Rate	1.9913	0.024	84.269	0.000	1.945	2.038
Body_Temp	-16.8434	0.353	-47.652	0.000	-17.536	-16.151

- **p-values** for all coeffs here are extremely small <<<0.05, indicating that ['Duration', 'Heart_Rate', 'Body_Temp'] are significant predictors of Calories
- **r-squared = 0.946**, indicating that the linear model including Duration, Heart_Rate, Body_Temp as predictors explains **94.6%** of the variation in Calories

Equation of line:

$$Calories = 470.35 + 6.641 \times Duration + 1.991 \times HeartRate - 16.843 \times BodyTemp$$

(Model 2) – Linear model with all features

Calories ~ Duration, Heart_Rate, Body_Temp, Height, Weight, Age, Gender

- repeat the same process but use all predictors

Results:

model_number	RMSE	r-squared
null	61.86	
1	14.57	0.946
2	11.29	0.967

- Model 2 including all features
    - Lower RMSE compared to null model and Model 1

p-values and r-squared:

OLS Regression Results

Dep. Variable:	Calories	R-squared:	0.967
Model:	OLS	Adj. R-squared:	0.967
Method:	Least Squares	F-statistic:	6.316e+04
Date:	Tue, 13 Feb 2024	Prob (F-statistic):	0.00
Time:	02:16:40	Log-Likelihood:	-57671.
No. Observations:	15000	AIC:	1.154e+05
Df Residuals:	14992	BIC:	1.154e+05
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	464.5795	11.102	41.847	0.000	442.819	486.340
Duration	6.6403	0.032	210.643	0.000	6.579	6.702
Heart_Rate	1.9903	0.018	107.785	0.000	1.954	2.027
Body_Temp	-16.9815	0.276	-61.484	0.000	-17.523	-16.440
Height	-0.1831	0.024	-7.481	0.000	-0.231	-0.135
Weight	0.3010	0.027	11.342	0.000	0.249	0.353
Age	0.5009	0.006	86.809	0.000	0.490	0.512
Gender	-1.2679	0.310	-4.087	0.000	-1.876	-0.660

- **p-values** for all coeffs here are extremely small <<<0.05, indicating that ['Duration', 'Heart_Rate', 'Body_Temp', 'Height', 'Weight', 'Age', 'Gender'] are significant predictors of Calories
- **r-squared = 0.975**, indicating that the linear model including Duration, Heart_Rate, Body_Temp as predictors explains 97.5% of the variation in Calories
- **Note:** the coeffs for Duration, Heart_Rate, and Body_Temp are not greatly changed from Model 1, indicating that these are by far the greatest contributors to the model.
    - Height, Weight, Age, Gender only have minor influence on Calories

Equation of line:

$$Calories = 463.57 + 6.640 \times Duration + 1.990 \times HeartRate - 16.982 \times BodyTemp - 0.183 \times Height + 0.301 \times Weight + 0.501 \times Age - 1.268 \times Gender$$

(Model 3) – Polynomial model with all features, and interaction between main predictors

Calories ~ Duration, Heart_Rate, Body_Temp, Height, Weight, Age, Gender

- Looking at curved shape of data in scatter matrix we saw:
  - Calories~Duration looks quadratic
  - Calories~Body_Temp looks quadratic
  - Calories~Heart_Rate looks cubic
- Using **sklearn (PolynomialFeatures degree=3),** build a model with Duration, Heart_Rate, Body_Temp terms (linear → cubic) + all combinations of interaction between these terms. Height, Weight, Age, Gender will stay as linear terms. Assess model performance against null and other models.
- Repeat process using **statsmodels** to get p-values and r-squared. (The statsmodels code is much more compact than sklearn)

Results:

model_number	RMSE	r-squared
null	61.86	
1	14.57	0.946
2	11.29	0.967
3	8.21	

- RMSE is a further improvement on the previous models
- **Problem:** I note that the sklearn and statsmodels outputs for this model are different
  - This is most likely due to the train-test split that is required to produce the sklearn output. By contrast, the statsmodels regression is trained on the entire dataset. I am not sure why, but differences in these methods did not produce a discrepancy in the results for our other models.
  - The statsmodels output is similar to sklearn, but the coefficients are slightly different. I'll include both the sklearn output and the statsmodels output separately below
  - The sklearn model is the predictive model we want to consider, so I will only use RMSE to assess it's performance.

Sklearn model coefficients:

I will not write the equation for brevity. The equation can be inferred from the table *(right) [more details can be found in corresponding jupyter notebook section]*:

Term	Coeff
Intercept	58058.67077
Duration	494.2601214
Heart_Rate	65.93890501
Body_Temp	-4630.655015
Duration^2	0.861890428
Duration*Heart_Rate	0.305197875
Duration*Body_Temp	-25.81705582
Heart_Rate^2	-0.207564439
Heart_Rate*Body_Temp	-2.493708847
Body_Temp^2	121.9341689
Duration^3	-0.000156231
Duration^2	0.000168001
Duration^2*Body_Temp	-0.021323592
Duration*Heart_Rate^2	-0.000441785
Duration*Heart_Rate*Body_Temp	-0.002635485
Duration*Body_Temp^2	0.330006237
Heart_Rate^3	0.000132234
Heart_Rate^2*Body_Temp	0.004423207
Heart_Rate*Body_Temp^2	0.022600493
Body_Temp^3	-1.062405974
Height	-0.197387327
Weight	0.313585191
Age	0.493846064
Gender	-1.267649976

- Note that many of these coefficients have a very small value, indicating that they will have a very small contribution to the model, even if the value of the predictor is large.
- We calculated the RMSE for model 3 using the train test split method. Normally, this should safeguard against overfitting - i.e. a low RMSE should suggest that the model, even if complex, should still perform well on out of sample data. Given that this data set is curated and does not have

the level of variance that we would normally see: it is possible that we have overfitted even though we used train test split. In the next step, we will try and simplify the model by removing some of the terms that have little impact on Calories.

Statsmodels output: p-values and r-squared (different coefficients from sklearn) – This model not used but included for reference

OLS Regression Results

Dep. Variable:	Calories	R-squared:	0.983
Model:	OLS	Adj. R-squared:	0.983
Method:	Least Squares	F-statistic:	3.732e+04
Date:	Tue, 13 Feb 2024	Prob (F-statistic):	0.00
Time:	13:47:33	Log-Likelihood:	-52806.
No. Observations:	15000	AIC:	1.057e+05
Df Residuals:	14976	BIC:	1.058e+05
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.722e+04	4.64e+04	1.664	0.096	-1.37e+04	1.68e+05
Duration	603.3603	411.687	1.466	0.143	-203.596	1410.316
Heart_Rate	64.8346	131.783	0.492	0.623	-193.476	323.145
Body_Temp	-6128.1796	3575.881	-1.714	0.087	-1.31e+04	880.984
Duration^2	1.1564	1.271	0.910	0.363	-1.335	3.648
Duration Heart_Rate	0.4299	0.849	0.506	0.613	-1.235	2.094
Duration Body_Temp	-31.8995	20.804	-1.533	0.125	-72.677	8.878
Heart_Rate^2	-0.3199	0.234	-1.369	0.171	-0.778	0.138
Heart_Rate Body_Temp	-1.9319	6.706	-0.288	0.773	-15.076	11.213
Body_Temp^2	160.3844	92.416	1.735	0.083	-20.762	341.531
Duration^3	0.0003	0.001	0.210	0.834	-0.002	0.003
Duration^2 Heart_Rate	0.0002	0.002	0.116	0.907	-0.003	0.003
Duration^2 Body_Temp	-0.0294	0.032	-0.928	0.354	-0.091	0.033
Duration Heart_Rate^2	-0.0006	0.001	-0.677	0.498	-0.002	0.001
Duration Heart_Rate Body_Temp	-0.0050	0.021	-0.239	0.811	-0.046	0.036
Duration Body_Temp^2	0.4131	0.264	1.562	0.118	-0.105	0.932
Heart_Rate^3	9.825e-05	0.000	0.423	0.672	-0.000	0.001
Heart_Rate^2 Body_Temp	0.0075	0.006	1.268	0.205	-0.004	0.019
Heart_Rate Body_Temp^2	0.0085	0.087	0.098	0.922	-0.162	0.179
Body_Temp^3	-1.3865	0.801	-1.731	0.083	-2.956	0.183
Height	-0.2089	0.018	-11.792	0.000	-0.244	-0.174
Weight	0.3301	0.019	17.186	0.000	0.292	0.368
Age	0.4966	0.004	118.877	0.000	0.488	0.505
Gender	-1.3827	0.224	-6.159	0.000	-1.823	-0.943

- This is similar but not the same as the sklearn model – the coefficients are different.
- Also note here that while the r-squared = 0.983, many of the predictors have a p-value> 0.005 (they appear not to be significant predictors). Some of these terms could be dropped whilst still preserving the predictive power of the model.

(Model 4) – Simplified linear regression with limited interaction

Calories ~(Heart_rate)^3, (Duration)^2, (Body_Temp)^2, Height, Weight, Age, Gender, Height*Weight

- Multiple methods for simplifying down from Model 3 were attempted
    - **Lasso Regression**
        - Unsuccessful → error message : ConvergenceWarning: Objective did not converge.
    - **Backward Stepwise Regression**
        - Starting with Upper model: Model 3 (poly, interactions, all predictors) → Lower model: Model 2 (linear, all predictors)
        - Stepping backwards by removing terms and checking AIC with each step to see if this improves the model. Stop when model is no longer improved.
        - Unsuccessful → algorithm a number of iterations but the end model was still too complex. *[see jupyter notebook]*
    - **Manual method was finally used:**
        - Recall the observations of relationships with Calories in scatter matrix:
            - Calories~Duration looks quadratic
            - Calories~Body_Temp looks quadratic
            - Calories~Heart_Rate looks cubic
        - Recall from correlation matrix
            - Height and Weight are highly correlated with one another
        - Therefore include the following polynomial terms
            - (Heart_rate)^3, (Duration)^2, (Body_Temp)^2
        - Include a single interaction term:
            - Height*Weight
        - Leave the rest of the predictors as linear terms
        - During the process , Gender coeff had p-value=0.735 >>0.05
            - Gender was removed from the final model
        - Successful → produce simplified polynomial model with very little loss of performance from Model 3.
            - **This is the preferred model out of the four.**

Final Results: Preferred model

model_number	RMSE	r-squared
null	61.86	
1	14.57	0.946
2	11.29	0.967
3	8.21	
4	8.92	0.980

- Based on RMSE and r-squared, there is very little loss of performance from Model 3.
- Model 4 is simpler and less likely to be overfitted to the sample data.

p-values and r-squared:

OLS Regression Results

Dep. Variable:	Calories	R-squared:	0.980
Model:	OLS	Adj. R-squared:	0.980
Method:	Least Squares	F-statistic:	1.036e+05
Date:	Tue, 13 Feb 2024	Prob (F-statistic):	0.00
Time:	16:41:38	Log-Likelihood:	-54055.
No. Observations:	15000	AIC:	1.081e+05
Df Residuals:	14992	BIC:	1.082e+05
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-54.9368	5.128	-10.713	0.000	-64.988	-44.885
Heart_Rate^3	7.83e-05	5.11e-07	153.170	0.000	7.73e-05	7.93e-05
Duration^2	0.1442	0.001	249.191	0.000	0.143	0.145
Body_Temp^2	0.0791	0.002	38.384	0.000	0.075	0.083
Height	-0.7427	0.026	-28.249	0.000	-0.794	-0.691
Weight	-1.5321	0.061	-25.243	0.000	-1.651	-1.413
Age	0.5075	0.004	113.693	0.000	0.499	0.516
Height_Weight_Interaction	0.0094	0.000	30.783	0.000	0.009	0.010

Equation of line:

$$Calories = -54.9368 + 7.7773 \times 10^{-5} \times HeartRate^3 + 1.4446 \times Duration^2 + 0.079416 \times BodyTemp^2 - 0.7427 \times Height - 1.5321 \times Weight + 0.0092282 \times Height \times Weight$$

4. General conclusions and recommendation

General remarks:

Dataset 1: Fitbit

- Initial analysis confirmed that TotalSteps, ActiveDistance, and ActiveMinutes metrics correlated positively with daily Calorie burn, as expected. Higher activity = more energy expenditure.
- People spent most of their day (81%) sedentary, with more vigorous activities forming smaller share of time during the day. Fitbit users undertook more light exercise (15.8%) than vigorous exercise (VeryActive = 1.7%, FairlyActive = 1.1%). Mean total steps per day was ~7600 - below the recommended 10000, for improved fitness. While the Fitbit app includes activity reminders, potentially the Fitbit could be doing more to promote movement during the day. Distribution of SedentaryMinutes had a lower standard deviation than ActiveMinutes - values bunching up around the mean (around 16.5 hours sedentary time) suggest that people like to get a certain amount of sedentary time per day without much variation.

- Basal Metabolic Rate (BMR) is the energy the body uses to maintain basic functions while at rest. Higher BMR is considered to be a measure of physical fitness. This was estimated using a simple linear regression of Calories on TotalSteps. BMR was the intercept of the fit line. This demonstrates how many features innate to an individual participant may be inferred just from activity metrics, even when demographic information is not available.
- While a lot of participants were diligent with using their wearable to log activity, logging a large proportion of 100% days, nobody logged activity for all minutes of the month. This is understandable since the wearable needs to be taken off to charge, isn't generally worn to shower or for prolonged water-based activities. The average % minutes logged per person across the month was 77%. This is a good result, indicating that Fitbit owners are committed to using their device. Pattern of logging across individuals suggests that motivation was a big factor contributing to logging frequency. Users that logged partial days ended up logging a lot of partial days in total, whereas there was a group of users that were very committed to logging full days. The motivation to log full days seemed to come and go in blocks. There would be a run of partially logged days, then a block of 1 or 2 full days, then another block of partial days. Potentially the Fitbit app could address these motivational issues more effectively through reminders/ incentives to wear the device. On device reminders can only be viewed with the Fitbit is worn, but perhaps more emphasis could be put on phone reminders for less motivated users.
- Nevertheless, increased frequency of logging activity/ sleep metrics did not correlate to increased daily calories burned, nor did it correlate to improved health metrics such as BMR. People the that logged more total minutes of metrics during the month were mostly logging sedentary minutes. They were just logging more time while they were inactive. More research needs to be done on how to encourage movement when the Fitbit is worn.
- A strong correlation between logging frequency and Calories burned, or logging frequency and BMR, was not observed. Consequently, logging metrics could not be utilized as a predictive basis for Calories burned as initially anticipated.
- In fact, most of the features strongly correlated with calories were auto-correlated by an existing mathematical relationship. For example, Steps and ActiveDistance metrics are used by the Fitbit algorithm to calculate daily Calories burned. Hence, these metrics correlate with calories in the analysis, but there is no point using them in a new model to predict Calories - it doesn't reveal anything new.
- There was a strong negative correlation between total time spent asleep and sedentary minutes during the day. Individuals that slept better were less sedentary during the day. However, most people were getting slightly less that the recommended 7 hours of sleep. On average, individuals slept more on a Sunday, but otherwise slept relatively consistently during the week.

- There were otherwise no strong patterns in activity/ sleep across the week. Highest Mean TotalSteps was on Saturday, but low on Sunday. Overall, Mean TotalSteps was higher on an average weekday than on the weekend.
- The most active hours of the day: Activity (mean steps/ hour) increased from 6am, reaching peaks around lunchtime (from 12-2pm ~540 steps) and after work (5-7pm 550-

599 steps), before dropping sharply at 8pm and continuing to decrease. There was a small amount of activity after midnight that may be attributed to restlessness/ getting out of bed. The lowest activity was at 3am.

Dataset 2: Exercise Lab

- The most important metrics for predicting calories using the Exercise Lab data were Duration of exercise (mins), skin surface Temperature (degrees Celsius), average Heart Rate during exercise session(bpm). The other predictors (Age, Gender, Height, Weight) had significantly less impact on calculating Calories burned in an individual exercise session.

- Ultimately a polynomial model was favoured, including all predictors but focusing on the main trio (Duration, Body_Temp, HR). Interaction between highly correlated variables Height and Weight was accounted for.

- Exercise duration, heart rate, and skin surface temperature can all be tracked using the Fitbit app and wearable. This result favours more recent Fitbit devices for accurate calorie tracking per exercise session. Fitbit Sense 2, Versa 4 and Charge 6 have a more sophisticated on-wrist temperature sensor, offering more detailed continuous tracking. The lower end devices such as the Fitbit Inspire 3 offer more limited temperature tracking. It is surprising that demographic information (Age, Gender, BMI) did not factor more into the Calories burned calculation. Perhaps in the future, even the more basic devices could offer advanced Temperature sensing in addition to HR for a more accurate recording of real-time energy expenditure.

Business Recommendations:

1. Enhance User Engagement and Motivation

- To address the high levels of sedentary behaviour, Fitbit could introduce more personalized, dynamic reminders and motivational challenges that encourage users to move more throughout the day. This could be enhanced through gamification or community challenges that leverage social motivation.
- Implement features that reward users for consistent logging and achieving daily and weekly activity goals, potentially integrating with wellness programs that offer real-world incentives.

2. Sleep and Activity Insight Integration

- Given the correlation between sleep and activity levels, Fitbit could develop more integrated insights that help users understand how improving one could beneficially impact the other. This could involve personalized advice or programs that aim to enhance both sleep quality and daily activity.

3. Advanced Feature Development

- Considering the importance of exercise duration, heart rate, and skin temperature in predicting calorie burn, future iterations of Fitbit devices should continue to refine these sensors and algorithms for even more accurate tracking.

- Invest in enhancing skin temperature tracking capabilities across all device tiers to make this feature a standard offering, given its relevance to overall health insights and calorie burn estimation.

4. Research and Development Focus

- Ongoing research into user behaviour patterns and health outcomes can inform the development of new features or improvements to existing ones. This includes exploring ways to encourage more active and less sedentary lifestyles among users.
- Collaboration with health professionals and institutions to validate the effectiveness of Fitbit's tracking capabilities and health insight recommendations.

Limitations and Future Data Collection Strategies

Dataset 1 analysis was severely limited by cohort size (33) and missing demographic data. It was disappointing that the weight_log and heart rate components of this data were so incomplete, as these would have provided a stronger basis to explore the relationships of metrics with general health. BMI could have been an interesting response variable to explore and HR would have been a good predictor for Calories. It would have been interesting to study these metrics in granular data over the course of a month.

Dataset 2 was limited by lack of documentation. While demographic data on participants was available, there was not much information on how the Exercise Lab metrics were collected from participants: e.g. types of activity, lab environment, exact date of collection. This makes it difficult to generalise conclusion to different types of exercise (anaerobic vs aerobic). Given that the data was curated for education purposes, there is some uncertainty about how the model will perform on out of sample data, given the variability of unprocessed real-world data.

In future, I would opt for larger scale Fitbit studies such as LifeSnaps for more comprehensive data collection. https://www.nature.com/articles/s41597-022-01764-x

Keep in mind that Fitbit has been acquired by Google, and thus is expected to leverage Google's resources to innovate and possibly increase its market presence.

By addressing these areas, Fitbit can improve its product offerings, engage users more effectively, and enhance its position as a leader in personalized health and fitness tracking.

Appendix A: R code for Pairs Plot

```
source("Rfunctions.R")

#Limit results to 3 dp
options(digits = 3, show.signif.stars = F)

#Read in data from cereal.csv and store in object
activity_sleep.df <- read.csv("activity_sleep_daily.csv", header = TRUE)
library(GGally)

## Loading required package: ggplot2

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

library(ggplot2)

# Define the columns  to include in the pairs plot
columns_to_include <- c('TotalMinutesAsleep', 'TotalTimeInBed', 'Calories'
, 'TotalSteps',
                        'VeryActiveMinutes', 'FairlyActiveMinutes', 'Light
lyActiveMinutes',
                        'SedentaryMinutes')

# Create the pairs plot with the specified columns
# Create the pairs plot with the specified columns
ggpairs(activity_sleep.df, columns = columns_to_include, cardinality_thres
hold = 30)
```