# Project Draft

Jessica Reyes, Seth Hillis, Niamh Corrigan, Andi Mellyn

2025-12-02

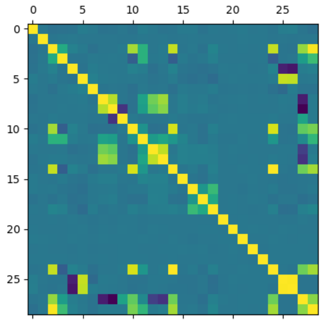
```
data <- read.csv("~/project-2/Data/new/data.csv")
```

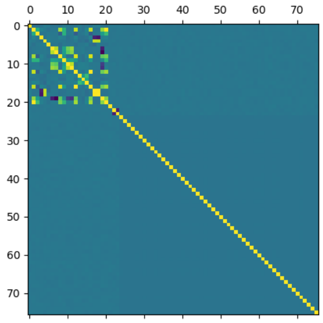
## Question/Problem

What features in our data have significant effects on calorie burns per 30 min, and does diet impact the calorie count during exercise?

## Data Preparation

The data required some preparation to make it read to be fit to a model. we started by determining the features that have the largest impact on the amount of calories burned. We started by making a correlation matrix to get an idea for what features are significant.





Here is a peak at the variable that correlated the most.

| Feature | Correlation with Burns_Calories_per_30_min |
|---|---|
| Burns_Calories_per_30_min | 1.000 |
| Sets | 0.465 |
| Reps | 0.343 |
| lean_mass_kg | 0.098 |
| Weight_kg | 0.091 |
| Height_m | 0.065 |
| BMI | 0.063 |
| Fat_Percentage | 0.054 |
| Experience_Level | 0.052 |
| Workout_Frequency_days_per_week | 0.044 |
| Resting_BPM | 0.038 |
| Water_Intake_liters | 0.026 |
| Workout_Pull.ups | 0.014 |
| Difficulty_Level_Beginner | 0.012 |

| Feature | Correlation with Burns_Calories_per_30_min |
|---|---|
| Workout_Crunches | 0.011 |

### Generalized Additive Model (GAM)

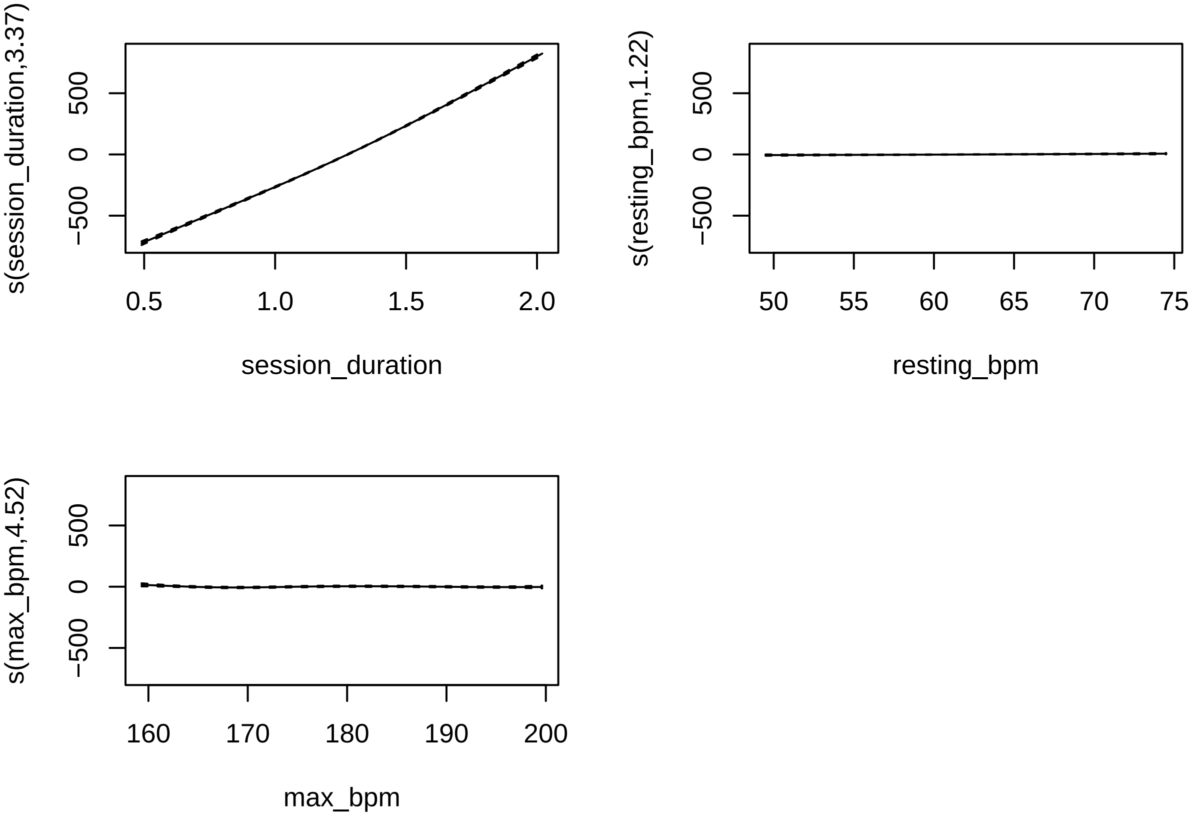
## Exercise Model

This model explains a substantial proportion of the variability in calories burned (84%), indicating that session duration, resting beats per minute, experience level, and workout type are important predictors.

```
exercise_gam <- mgcv::gam(
  calories_burned ~
    s(session_duration) +
    s(resting_bpm) +
    s(max_bpm) +
    experience_level +
    workout_type,
  data = data,
  method = "GCV.Cp"
)

summary(exercise_gam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## calories_burned ~ s(session_duration) + s(resting_bpm) + s(max_bpm) +
##     experience_level + workout_type
##
## Parametric coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1345.000      6.215   216.4   <2e-16 ***
## experience_level  112.255      3.171    35.4   <2e-16 ***
## workout_type     -179.155      1.265  -141.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                       edf Ref.df        F p-value
## s(session_duration) 3.367  4.178 5703.610  <2e-16 ***
## s(resting_bpm)      1.215  1.398    4.833  0.0294 *
## s(max_bpm)          4.525  5.549    1.960  0.0665 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.84   Deviance explained =   84%
## GCV =  40328  Scale est. = 40303     n = 20000
```

```
plot(exercise_gam, pages = 1)
```



## Smooth Terms form GAM

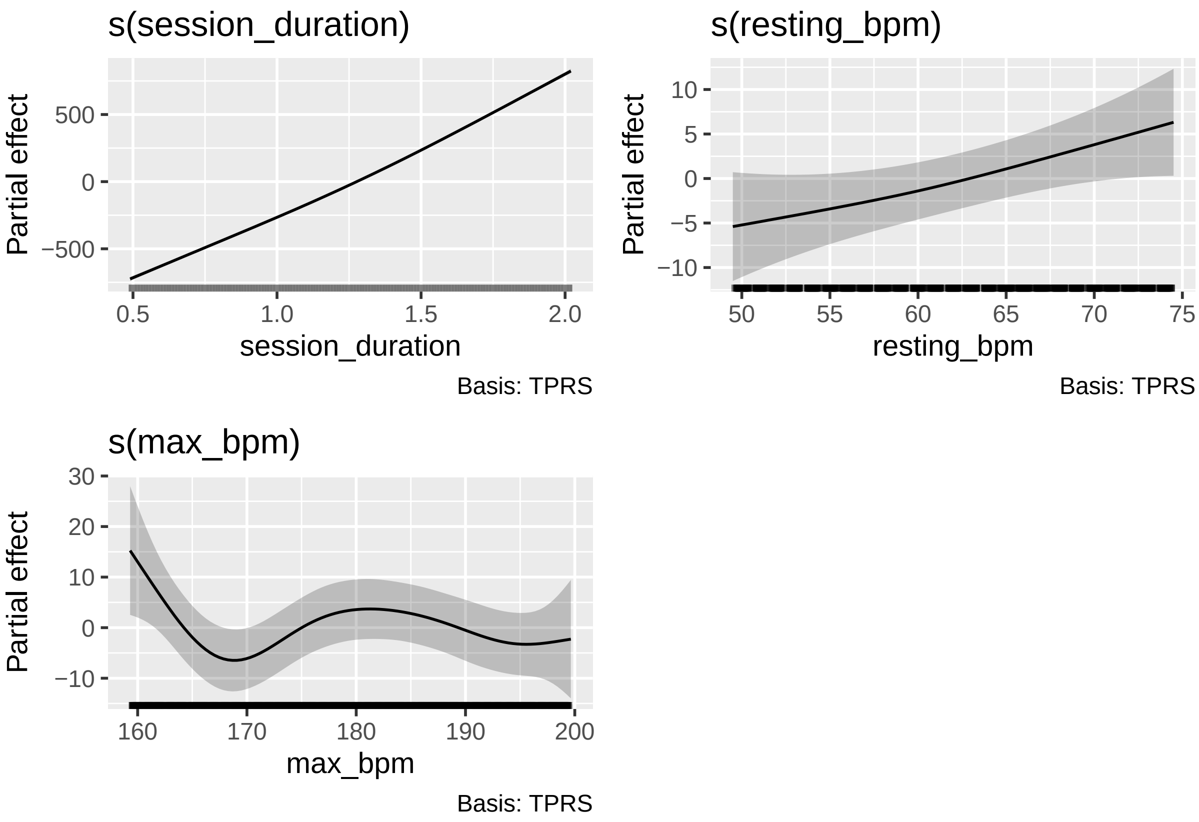
As expected, session duration has a strong, nearly linear positive effect on calories burned. Longer sessions consistently result in more calories burned, showing a positive relationship.

Resting BPM shows a moderate positive relationship with calories burned. As resting BPM increases, calories burned also tend to increase, though the effect is less pronounced than for session duration.

Maximum BPM demonstrates a complex non-linear effect with fluctuations across its range.

These patterns highlight the importance of considering non-linear relationships in understanding how phsiological and workout factors influence calories burned.

```r
draw(exercise_gam)
```



### Generalized Additive Model (GAM)

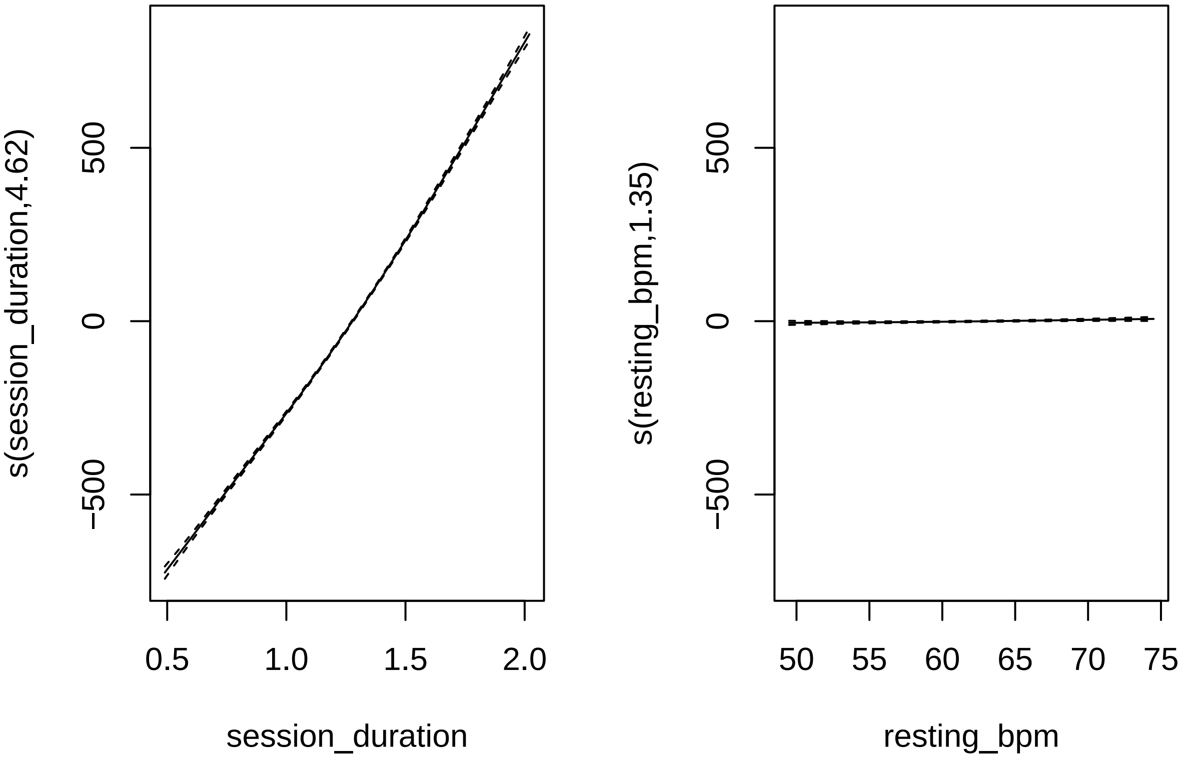
## Exercise & Diet Model

Including diet type in the model reveals a statistically significant but very small positive effect on calories burned. However, the overall model fit remains unchanged, suggesting that diet type does not substantially improve the model's ability to explain variability in calories burned beyond session duration, heart rate measures, experience level, and workout type.

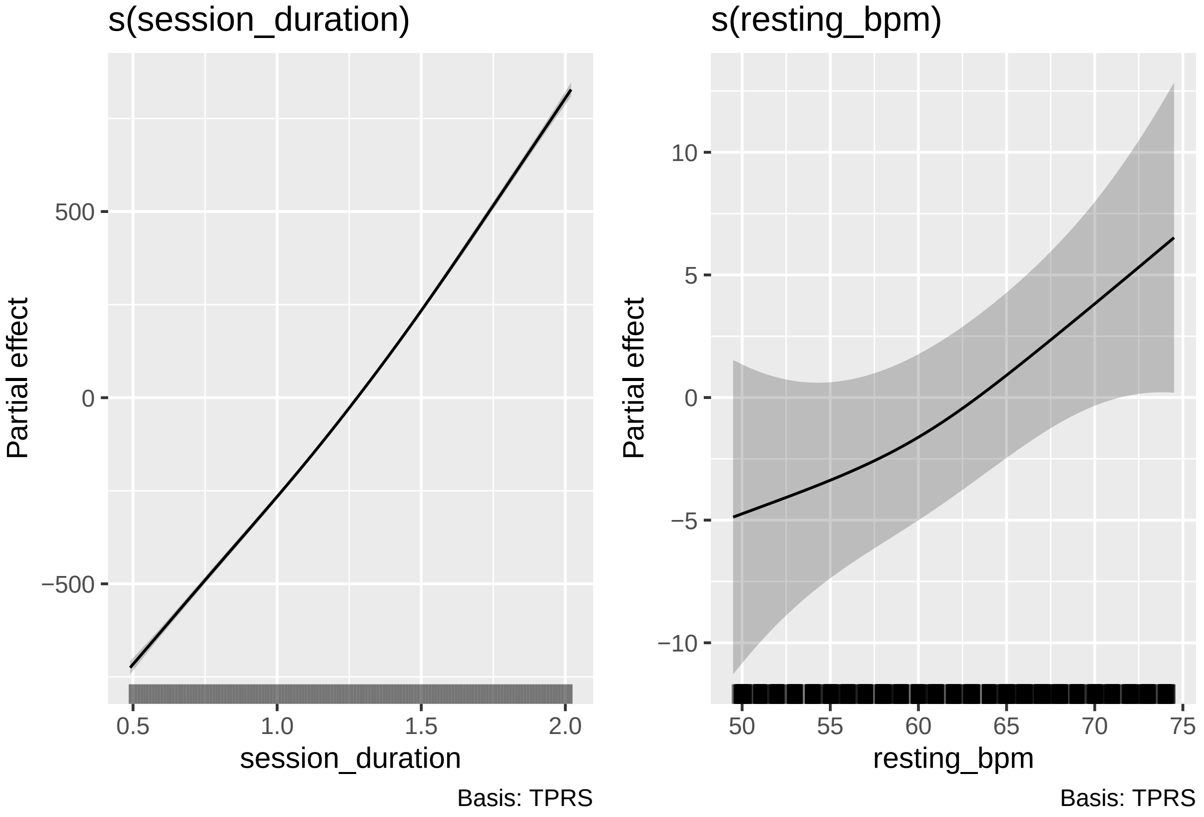
```r
combined_gam2 <- mgcv::gam(
  calories_burned ~
    s(session_duration) +
    #s(avg_bpm) +
    #s(max_bpm) +
    s(resting_bpm) +
    #s(weight) +
    #s(age) +
    #s(bmi) +
    experience_level +
    workout_type +
    #s(water_intake) +
    #s(workout_frequency) +
    #daily_meals_frequency +
    #s(carbs) +
    #s(proteins) +
    #s(fats) +
    diet_type,
  data = data,
  method = "REML"
)

summary(combined_gam2)
```

```
## 
## Family: gaussian 
## Link function: identity 
## 
## Formula:
## calories_burned ~ s(session_duration) + s(resting_bpm) + experience_level + 
##     workout_type + diet_type
## 
## Parametric coefficients:
##                   Estimate Std. Error  t value Pr(>|t|)    
## (Intercept)      1341.8659     6.7129  199.894   <2e-16 ***
## experience_level  111.6165     3.2879   33.947   <2e-16 ***
## workout_type     -179.1245     1.2650 -141.602   <2e-16 ***
## diet_type           1.7044     0.8357    2.039   0.0414 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Approximate significance of smooth terms:
##                       edf Ref.df        F p-value    
## s(session_duration) 4.623  5.675 4261.709  <2e-16 ***
## s(resting_bpm)      1.349  1.618    4.291  0.0401 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## R-sq.(adj) =   0.84   Deviance explained =   84%
## -REML = 1.3442e+05  Scale est. = 40316     n = 20000
```

```
plot(combined_gam2, pages = 1, se = TRUE)
```



```r
draw(combined_gam2)
```



## Classification Tree / Random Forest

The single decision tree with CP of .01lead to RMSE of 23.670 and explains 41.298% of variance explained, which by itself is not too useful.

//Add model parameters from .fit

//Add graphs here

When the tree are made into a random forest of 500 trees the RMSE drops to 9.844 and explains 89.847% of the variance in the data.

//add dendragram here