Statistical Analysis of Lifestyle Predictors on Resting Heart Rate

Angel D Rodas*

Resting heart rate (RHR) is a widely used biomarker of cardiovascular fitness and physiological stress, its relationship with demographic and lifestyle factors is complex and highly variable both between and within individuals. Using a dataset of 20,000 adults, we evaluated how Age, BMI, Water Intake, Gender, and Workout Frequency relate to RHR through three modeling strategies: (1) standard linear regression, (2) a negative inverse transformation of RHR to address diagnostic violations, and (3) a generalized additive model (GAM) with nonlinear smooth terms.

Both linear models explained minimal variance (Adjusted $R^2 \approx 0.002$) and showed limited practical predictive value. The inverse transformation improved residual behavior but did not enhance explanatory power. The GAM provided the strongest performance (Adjusted $R^2 = 0.0211; 2.25\%$ deviance explained), revealing significant nonlinear effects in our data for Age, BMI, and Water Intake on RHR (all $p < 2 \times 10^{-16}$). Gender effects were negligible, and Workout Frequency demonstrated only a small and statistically insignificant impact (P).

Across all models, effect sizes were small, suggesting that the measured lifestyle and demographic variables account for only a small portion of resting heart rate variability in the data. These findings support that RHR is influenced primarily by physiological, behavioral, or genetic factors not included in this dataset, underscoring the need for richer measurements in future studies.

I. Nomenclature

RHR	=	Resting Heart Rate
BPM	=	Beats Per Minute
BMI	=	Body Mass Index
H_0	=	Null Hypothesis
H_1	=	Alternative Hypothesis
ANOVA	=	Analysis of Variance
QQ Plot	=	Quantile-Quantile Plot
RMSE	=	Root Mean Squared Error
AIC	=	Akaike Information Criterion
BIC	=	Bayesian Information Criterion
GAM	=	Generalized Additive Model
Cook's D	=	Cook's Distance
DOF	=	Degrees of Freedom
edf	=	Effective Degrees of Freedom
Type I Error	=	False Positive (Rejection of a true null hypothesis)
Type II Error	=	False Negative (Failure to reject a false null hypothesis)

II. Introduction

A. Background

Resting heart rate (RHR), measured in beats per minute (BPM), is an individual's heart rate measured while at rest (sitting or lying down) and in a calm state. Similar to many types of biometric data, RHR exhibits high individual variability. [1]

*MS Student, UC Davis Mechanical and Aerospace Engineering

Resting heart rate is of particular interest because, in general, it serves as an indicator of cardiovascular health; additionally, resting heart rates outside the normal range can indicate underlying health conditions. Higher heart rates have been linked to a higher likelihood of developing cardiovascular disease. Conversely, lower RHRs have been associated with greater cardiovascular fitness and overall health. [1–5]

B. Motivation

Determining which predictors are the most statistically significant for predicting RHR may explain which lifestyle, demographic, and body composition factors provide better cardiovascular health. Investigating trends using statistical analysis can also provide insight into the dynamics of resting heart rate across different magnitudes of factors. These insights can be expanded to broader interpretations regarding RHR trends in response to changes in lifestyle, demographics, and body composition.

Identifying which predictors have the greatest influence on RHR can help provide recommendations for improving cardiovascular health from a data-driven perspective.

C. Objective

The aim of this project is to determine whether it is possible to predict the resting heart rate (RHR) of an individual given various factors such as age, gender, body composition, and lifestyle. Specifically:

Primary Question:

Which demographic, lifestyle, and body composition factors influence resting heart rate among adults, and how do they do so?

Hypotheses:

- H_0: Lifestyle factors have no significant relationship with resting heart rate.
- H_1: At least one lifestyle or body composition variable significantly predicts resting heart rate.

III. Methods

A. Dataset Description

This dataset was obtained on Kaggle, a large online platform and community for data science and machine learning enthusiasts. One of the many features of Kaggle is hosting a large number of publicly available datasets.

The specific dataset used for this analysis was the "Life Style Data" dataset provided by user Omar Essa [6].

The dataset originally contained biometric data, lifestyle data, and performance data from a single workout session, which varied in duration and type for each individual.

From the original 54 columns in the dataset, we extracted the following factors of interest for the analysis:

Table 1 Variables used in the analysis

Category	Variables	Type
Dependent Variable	Resting_BPM	Continuous
Demographic	Age, Gender	Age: Continuous; Gender: Categorical
Body Composition	BMI, Fat_Percentage	Continuous
Activity	Workout_Frequency (days/week),	Ordinal
Dietary and Hydration	Daily_Calories_In,Water_Intake (liters)	Continuous
Excluded	Everything_else	Mixed

After conducting a correlation analysis among the variables to minimize inter-predictor correlation and multi-collinearity (as shown in Figure 7), BMI, Fat Percentage, and Daily Calorie Intake were identified as having a high degree of correlation. Due to this strong correlation between these predictor variables, Fat Percentage and Daily Calorie Intake were removed from consideration in favor of BMI, a much more commonly used predictor in previous literature.

Table 2 presents the summary statistics for the remaining predictors under consideration.

Table 2 Descriptive statistics for all variables (n = 20,000).

Variable	Mean	SD	Median	Min	Max	SE
Resting BPM	62.20	7.29	62.20	49.49	74.50	0.05
Age	38.85	12.11	39.86	18.00	59.67	0.09
BMI	24.92	6.70	24.12	12.04	50.23	0.05
Water Intake (L/day)	2.63	0.60	2.61	1.46	3.73	0.00
Workout Frequency (days/week)	2.32	0.91	2.00	1.00	4.00	0.01

From these variables, Gender was treated as a factor given its categorical status. Since the data were not normally distributed, non-parametric tests were performed for comparison. ANOVA was also used; however, this was primarily to test if ANOVA methods would be robust to this dataset's departures from normality. The results of these tests are shown in Table 4.

Workout Frequency was more difficult to classify. We ran analyses treating it both as a categorical variable (binning 2, 3, 4, and 5 days/week) and as a numeric variable. Ultimately, Workout Frequency was classified as a factor for group comparison using the Wilcoxon test to ensure robustness, given the heavy tails observed in the QQ plot (Figure 6). However, Workout Frequency was treated as numeric for the linear regression model, given the seemingly monotonic and linear nature of its relationship with the response variable, as shown in Figure 9.

Given the distribution of Workout Frequency as a category, we used the Kruskal-Wallis test because the distributions in all categories deviated from normality and displayed heavy tails (Figure 6). ANOVA was also performed to investigate its robustness to these specific deviations from normality (Table5).

B. Model Building

Utilizing 10-fold cross-validation and all the variables included in Table 2, a four-parameter model including Age, BMI, Workout Frequency, and Water Intake performed best based on the Root Mean Squared Error (RMSE). A table summarizing the results of the cross-validation is provided below in Table 6.

From the cross-validation results, we observed that a four-parameter model minimized the error. Specifically, the model incorporating Age, BMI, Workout Frequency, and Water Intake was identified as the optimal subset.

Analysis of the diagnostic plots led to the development of two additional models to better fit the data:

1) A linear model with a negative inverse transformation applied to the dependent variable (Resting Heart Rate).
2) A Generalized Additive Model (GAM) incorporating all non-correlated predictors.

The final selected models are presented in Table 3.

While other GAM configurations were tested, this specific model yielded the best fit, AIC, and BIC values, leading to its selection as the final GAM model.

Table 3 Model Equations for Linear, Transformed Linear, and GAM Approaches

Model	Equation
Linear Regression	$\text{Resting\_BPM}_i = \beta_0 + \beta_1(\text{Age}_i) + \beta_2(\text{BMI}_i) + \beta_3(\text{WorkoutFrequency}_i) + \beta_4(\text{WaterIntake}_i) + \varepsilon_i$
Negative Inverse Transform Model	$\text{RHR\_inv}_i = -\frac{1}{\text{Resting\_BPM}_i}$ $\text{RHR\_inv}_i = \beta_0 + \beta_1(\text{Age}_i) + \beta_2(\text{BMI}_i) + \beta_3(\text{WaterIntake}_i) + \beta_4(\text{WorkoutFrequency}_i) + \varepsilon_i$
Generalized Additive Model (GAM)	$\text{Resting\_BPM}_i = \beta_0 + \beta_1(\text{Gender}_i) + \beta_2(\text{WorkoutFrequency}_i) + f_1(\text{Age}_i) + f_2(\text{BMI}_i) + f_3(\text{WaterIntake}_i) + \varepsilon_i$

C. Diagnostics

Examination of the diagnostic plots for the residuals revealed that the linear model showed clear signs of assumption violations. Figure 1 displays a clear deviation from normality, which is most evident in the Normal Q-Q plot. Additionally, the Cook's distance plot identifies several influential outlier points. There is also clear heteroscedasticity, most easily observed in the Scale-Location plot, which shows a distinct "bow" shape in the residuals.

When examining the diagnostic plots after applying the negative inverse transformation (Figure 2), the residuals still exhibit clear signs of deviation from normality. However, the influential outliers have been mitigated, as evident from the Cook's distance plot. The Scale-Location plot no longer displays the bow shape at the top, but rather at the bottom, indicating that heteroscedasticity is still present despite the transformation.

The residuals for the GAM demonstrated the best adherence to normality among the tested models. For reference, this GAM was fitted assuming a Gaussian distribution with an identity link function.

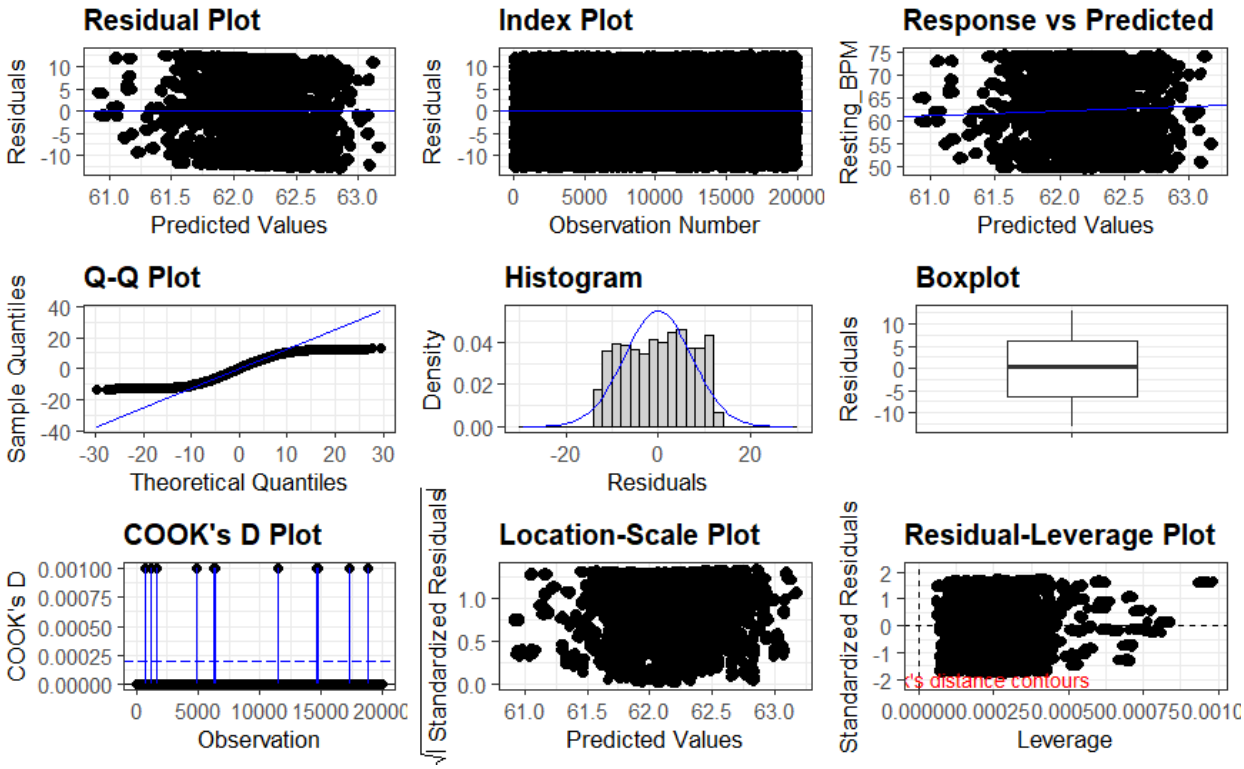


Fig. 1 Diagnostic plots for the linear regression model of Resting Heart Rate. Panels include Residuals vs Fitted, Index Plot, Response vs Fitted, Q-Q Plot, Histogram of Residuals, Boxplot of Residuals, Cook's Distance, Location-Scale Plot, and Residual-Leverage Plot.

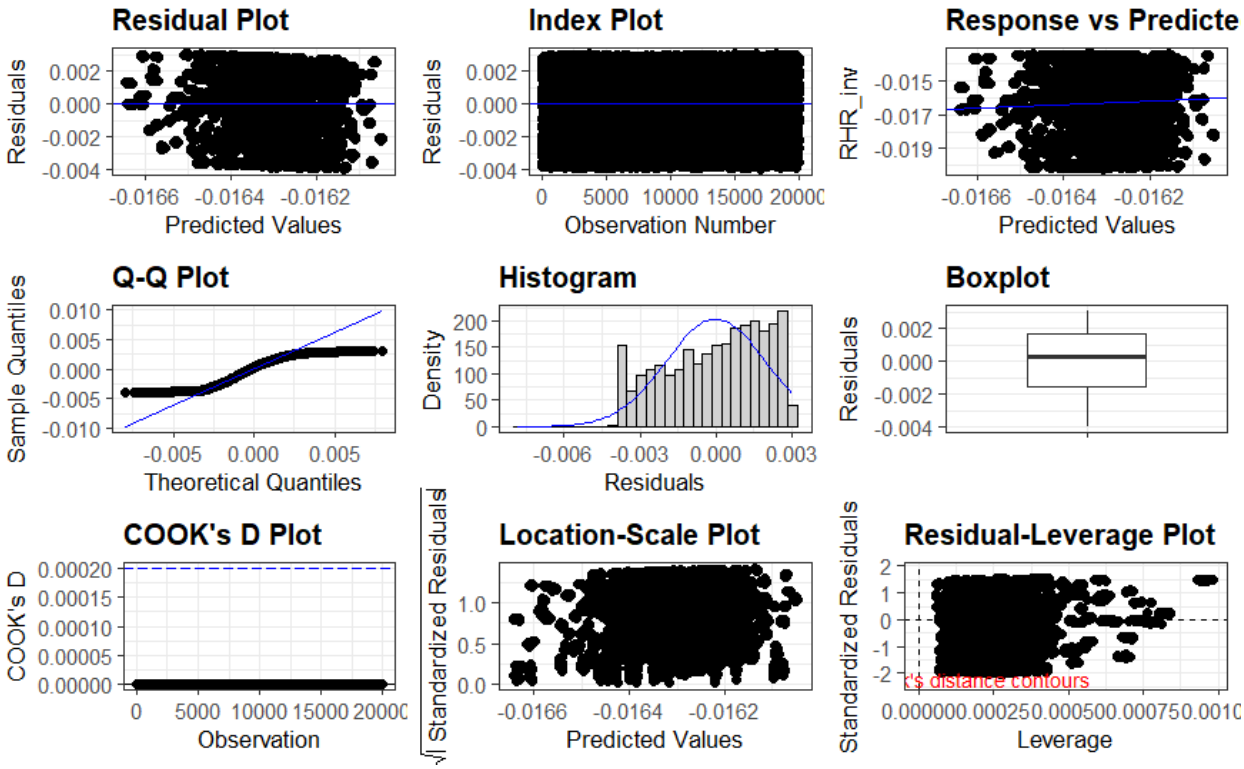


Fig. 2 Diagnostic plots for the linear regression model using the negative inverse transformation of Resting Heart Rate. Panels include Residuals vs Fitted, Index Plot, Response vs Predicted, Q–Q Plot, Histogram of Residuals, Boxplot, Cook's Distance, Location-Scale Plot, and Residual-Leverage Plot.

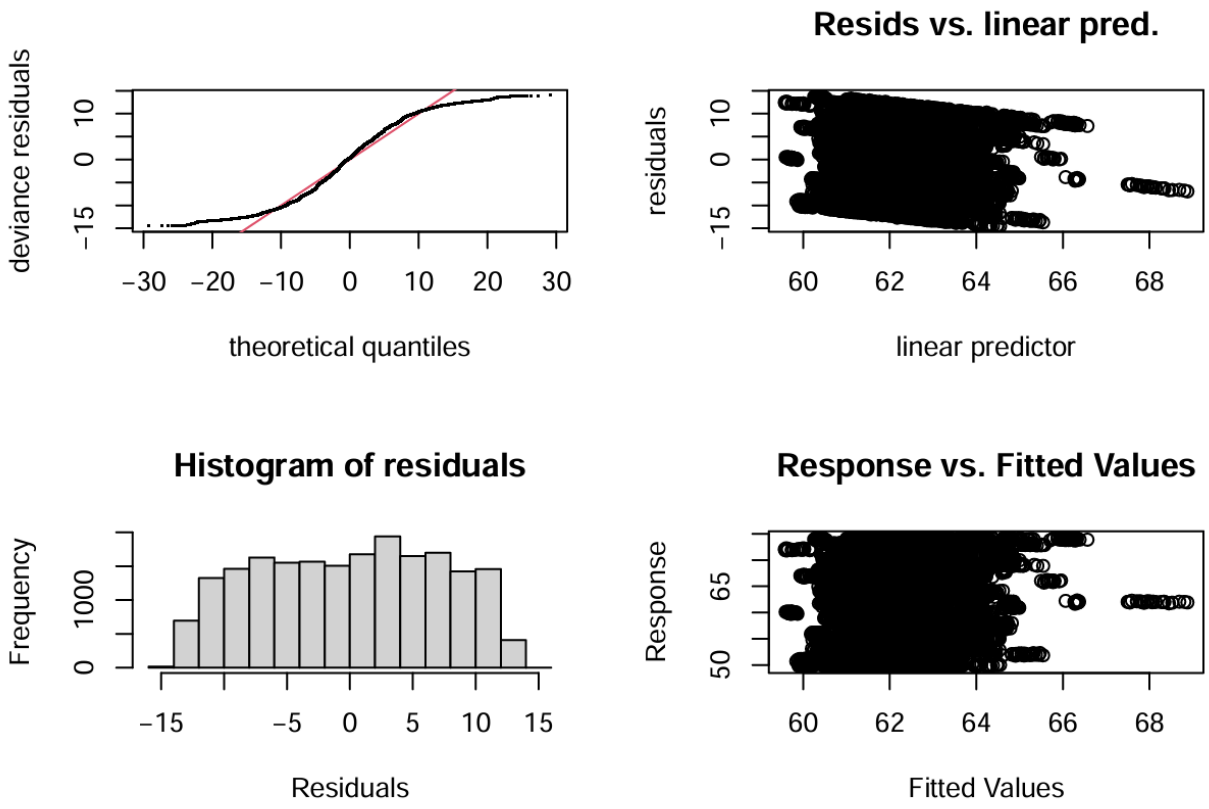


Fig. 3 Diagnostic plots for the Generalized Additive Model (GAM), including: (1) Q–Q plot of deviance residuals, (2) residuals vs. linear predictor, (3) histogram of residuals, and (4) response vs. fitted values.

IV. Results

A. Exploring the Data

1. Resting Heart Rate Distribution

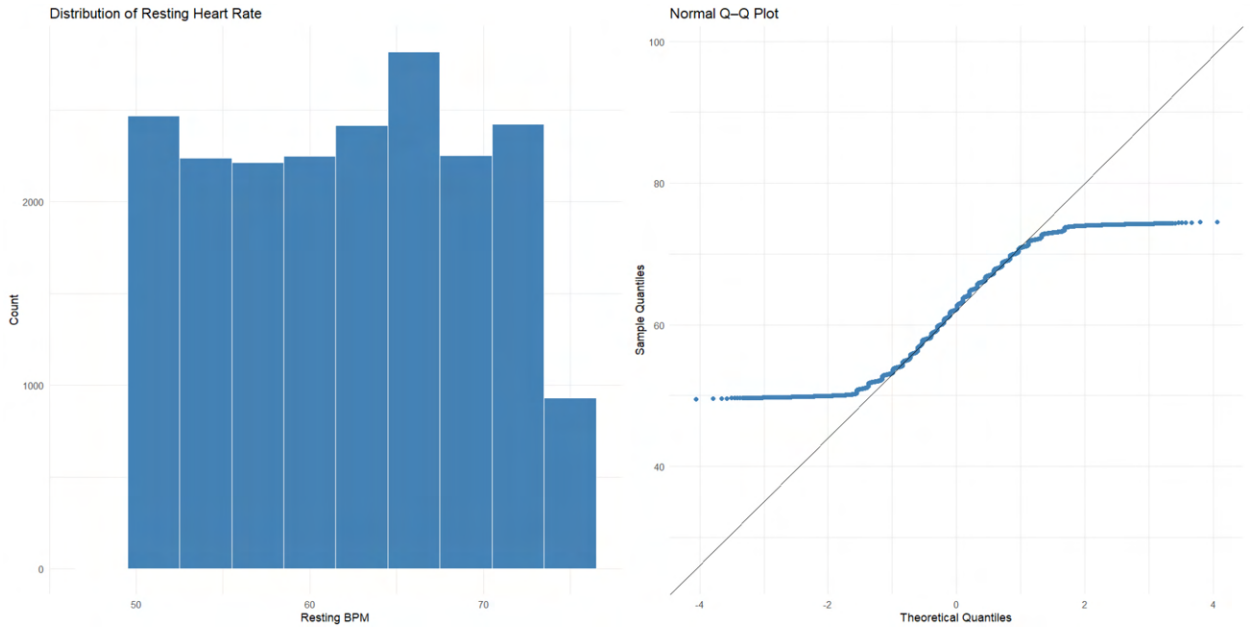


Fig. 4 Distribution and normality diagnostics for resting heart rate. Left: histogram of Resting BPM. Right: normal Q–Q plot of Resting BPM.

From the data visualization, it is evident that the distribution of RHR is not normal, exhibiting very pronounced tails. However, the central region of the distribution shows better adherence to normality.

2. Gender

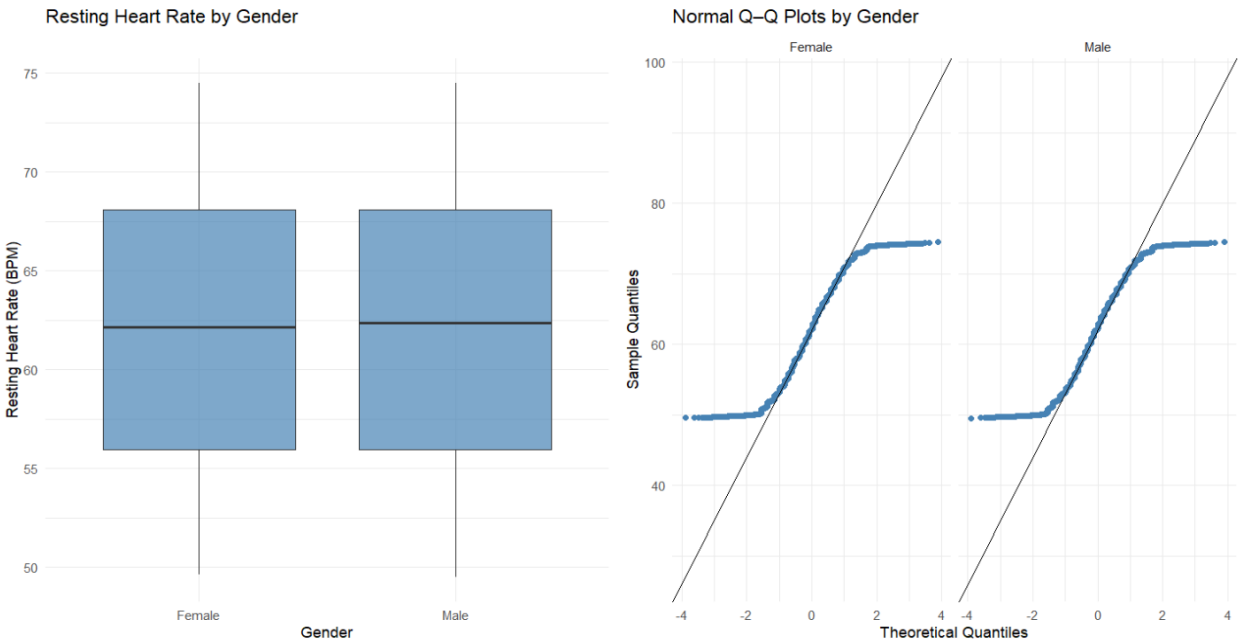


Fig. 5 Gender-based distribution and normality diagnostics for resting heart rate. Left: boxplot of Resting BPM by gender. Right: normal Q–Q plots for Female and Male subgroups.

Table 4 Summary of ANOVA, Wilcoxon rank-sum, and Levene's tests assessing gender differences in Resting BPM.

ANOVA Source	Df	Sum Sq	Mean Sq	F / p-value
Gender	1	8	8.46	$F = 0.159,\ p = 0.69$
Residuals	19998	1,062,631	53.14	–

Wilcoxon Rank-Sum Test	Statistic	p-value	
Resting BPM ~ Gender	$W = 49{,}841{,}984$	0.6994	–

Levene's Test (Median Centered)	Df	F value	p-value
Group	1	0.3823	0.5364
Residuals	19998	–	–

The QQ plots reveal deviations from normality, specifically heavy tails similar to those observed in the overall RHR distribution. In addition, the statistical tests indicate no significant difference between Male and Female RHRs.

3. Workout Frequency

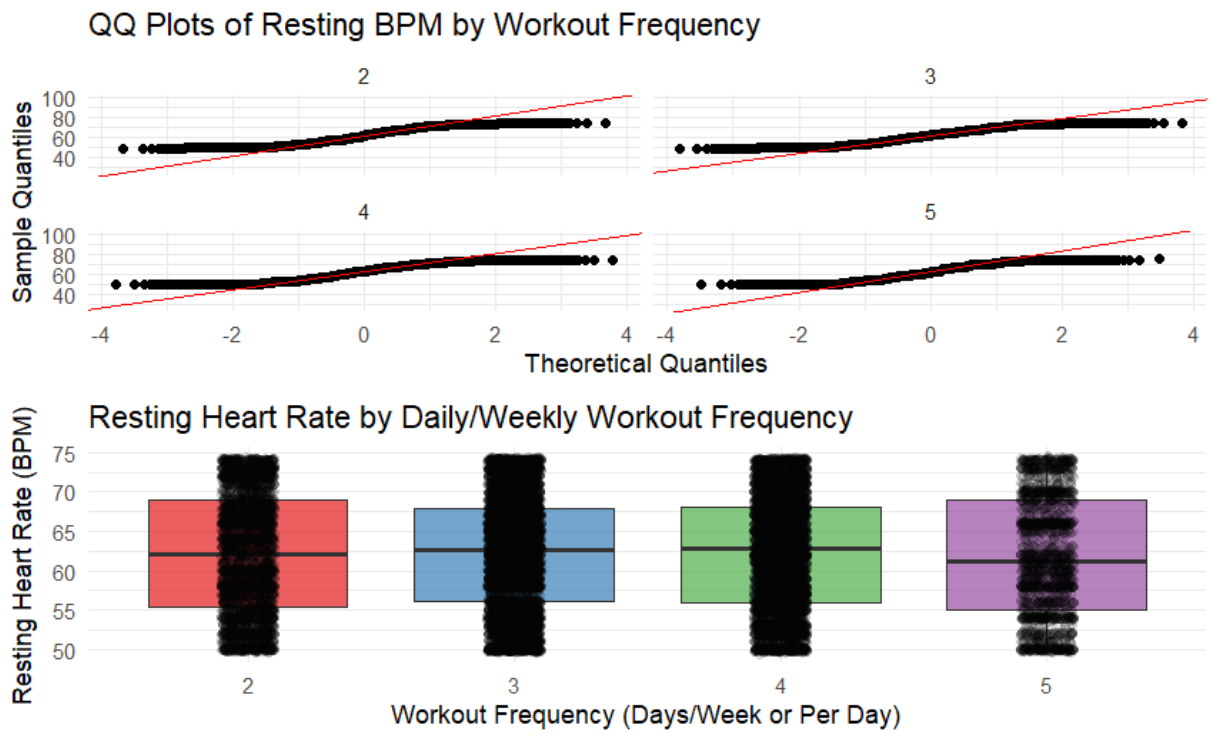


Fig. 6 QQ plots and boxplot comparison of Resting Heart Rate across Workout Frequency groups. The top panel shows normality diagnostics for each workout frequency, while the bottom panel visualizes the distribution of Resting BPM using a boxplot with jittered observations.

Workout Frequency Group	n
2 days/week	4050
3 days/week	7605
4 days/week	6266
5 days/week	2079

Levene's Test (Median Centered)	Df	F value	p-value
Workout_Frequency	3	0.3823	0.5364
Residuals	19998	–	–

Kruskal–Wallis Test	Chi-Sq	Df	p-value
Resting BPM ~ Workout Frequency	3.8968	3	0.2728

ANOVA Source	Df	Sum Sq	Mean Sq	F / p-value
Workout_Frequency	3	222	73.94	$F = 1.392,\ p = 0.243$
Residuals	19996	1,062,418	53.13	–

Table 5 Summary of sample sizes, ANOVA, Levene test, Tukey post-hoc comparisons, and Kruskal–Wallis test for Resting BPM across workout frequency groups.

The diagnostic plots relay that all Workout Frequency categories exhibit the same violations of normality observed in the Gender and RHR distributions. Consistent with these visual findings, the statistical tests detected no significant differences in the median or mean resting heart rates across workout frequency groups (Kruskal-Wallis test and ANOVA, respectively).

4. Post Correlation Matrix Analyses

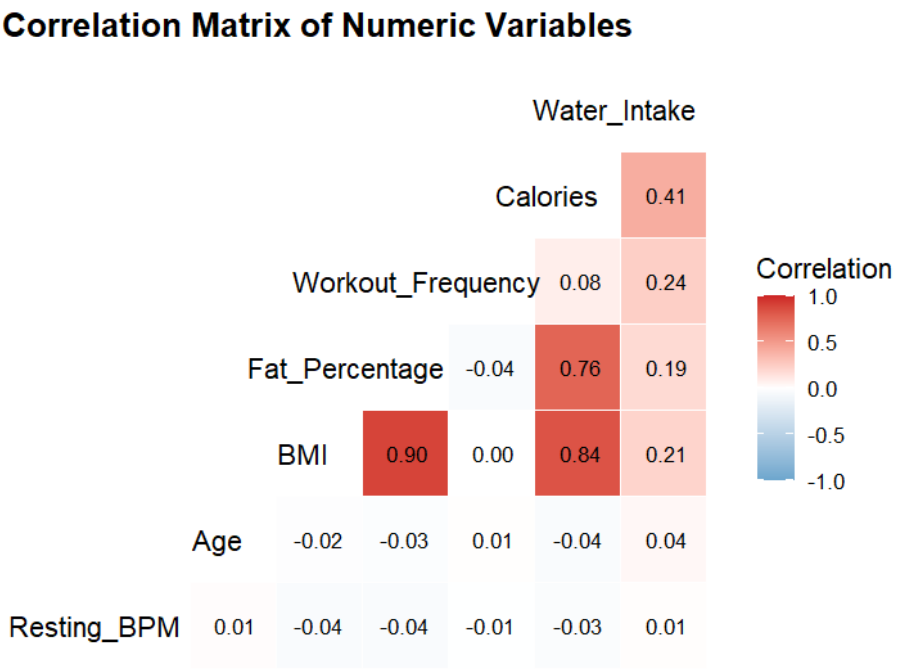


Fig. 7 Correlation Matrix of Numeric Variables

The correlation matrix reveals a cluster of strong correlations between BMI, Fat Percentage, and Daily Calorie Intake. Additionally, we observe a low degree of correlation between RHR and any single predictor.

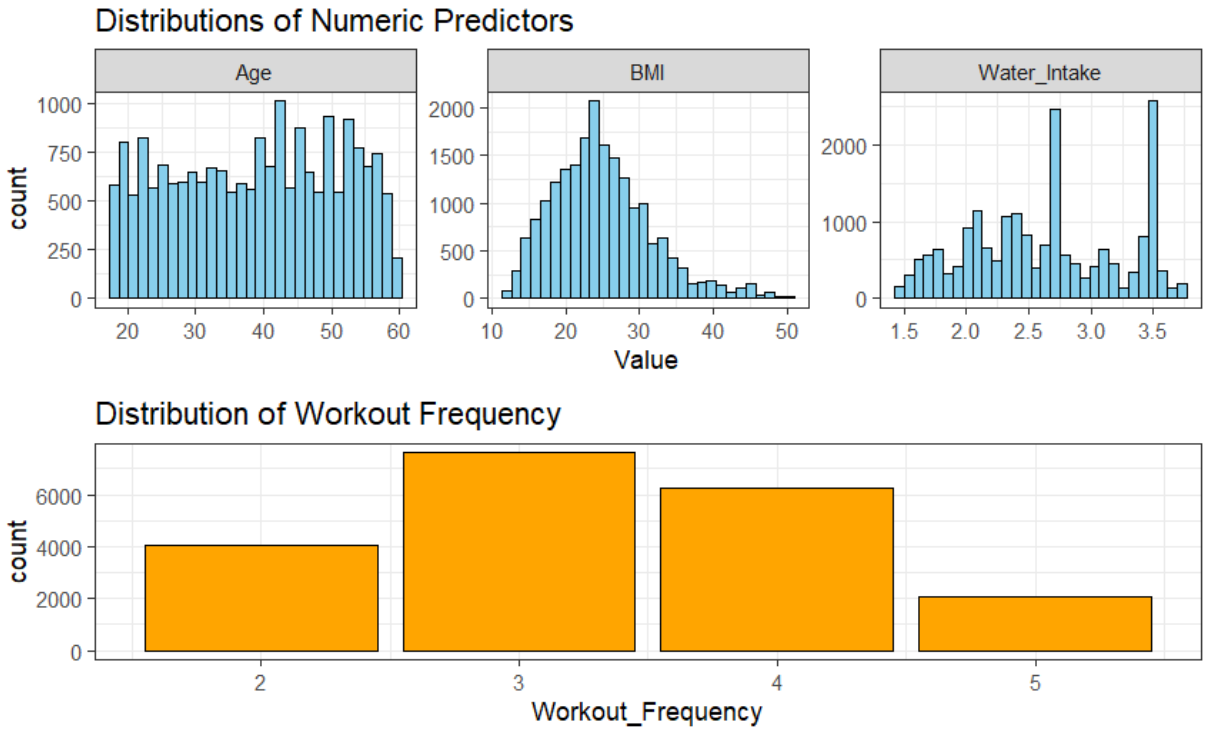


Fig. 8 Distributions of predictor variables used in the linear model: Age, BMI, Water_Intake, and Workout_Frequency.

The data reveals that Age is almost uniformly distributed, although, it exhibits peaks around 20 and 45 years. BMI shows a heavily skewed distribution concentrated around the 25 range. Water Intake also displays a concentration of values around 2.75 and 3.5 liters per day, while Workout Frequency follows an approximately normal distribution.

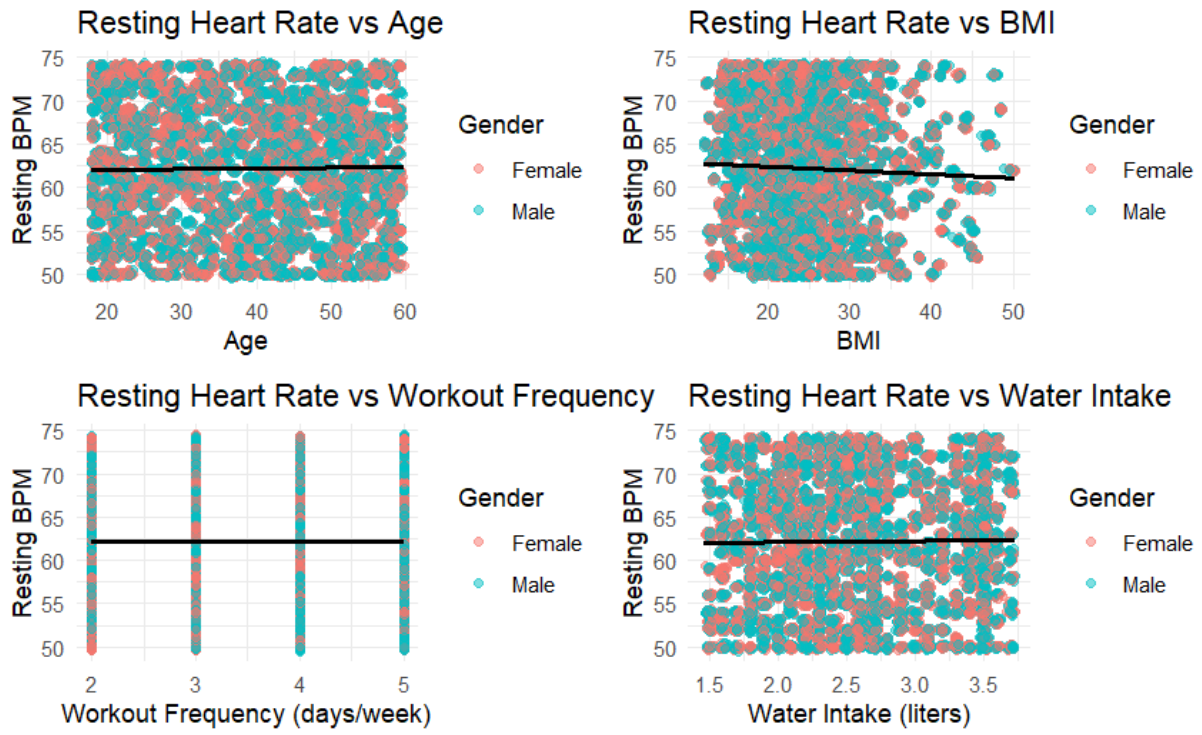


Fig. 9 Scatter plots of resting heart rate against Age, BMI, Workout Frequency, Gender, and Water Intake with fitted linear trends.

The plots of the non-correlated predictors show a very weak response to any single variable; however, the most prominent trend was observed in BMI.

B. Model Selection Results

Table 6 Linear Model Selection Performance Across Subset Sizes

nvmax	Predictors	RMSE	R^2	MAE	RMSE SD	R^2 SD	MAE SD
1	1	7.2834	0.00207	6.2799	0.03967	0.00155	0.04846
2	2	7.2820	0.00252	6.2794	0.04021	0.00220	0.04917
3	3	7.2843	0.00182	6.2802	0.03985	0.00147	0.04846
4	4	7.2816	0.00271	6.2773	0.04071	0.00234	0.04992
5	5	7.2820	0.00257	6.2780	0.04048	0.00224	0.04929
6	6	7.2820	0.00257	6.2780	0.04048	0.00224	0.04929

Table 7 Model Comparison: Linear Model, Inverse Linear Model, and GAM

AIC Comparison			
Model	df	AIC	
m.final	6.00000	136178.0	
gam_fit	31.43349	135820.6	
BIC Comparison			
Model	df	BIC	
m.final	6.00000	136225.5	
gam_fit	31.43349	136069.0	
ANOVA (LM vs GAM)			
Term	Res.Df	RSS	p-value
Linear Model	19995	1,060,141	—
GAM Model	19971	1,038,715	$< 2.2 \times 10^{-16}$ ***
Predictive Performance (Original BPM Scale)			
Model	RMSE	MAE	R^2
Linear	7.280593	6.276755	0.002352
Inverse Linear	7.332215	6.331437	-0.011846
GAM	7.206646	6.199122	0.022514

Based on the 10-fold cross-validation results, the four-parameter model provided the best average predictive performance across all candidates. Therefore, the model comprising Age, BMI, Workout Frequency (numeric), and Water Intake was selected.

C. Model Outputs

Table 8 Linear Regression Model for Resting Heart Rate

Predictor	**Estimate**	**Std. Error**	**t-value**	**p-value**
Intercept	62.7017	0.3486	179.866	< 2e-16 ***
Age	0.007103	0.004255	1.670	0.0950 .
BMI	-0.049601	0.007880	-6.294	3.15e-10 ***
Workout Frequency	-0.089121	0.058321	-1.528	0.1265
Water Intake	0.285476	0.090030	3.171	0.00152 **
Residual standard error: 7.282 (df = 19995)				
Multiple R^2: 0.002352 *Adjusted* R^2: 0.002152				
F-statistic: 11.78 on 4 and 19995 df *p-value*: 1.467e-09				

Table 9 Linear Regression Model Using Negative Inverse Transformation of Resting BPM

Predictor	Estimate	Std. Error	t-value	p-value	Std. Coef. (β^*)
Intercept	-1.617×10^{-2}	9.373×10^{-5}	-172.522	$< 2 \times 10^{-16}$ ***	–
Age	1.960×10^{-6}	1.144×10^{-6}	1.713	0.08672 .	0.0121
BMI	-1.297×10^{-5}	2.119×10^{-6}	-6.121	9.49×10^{-10} ***	-0.0443
Water_Intake	7.249×10^{-5}	2.421×10^{-5}	2.995	0.00275 **	0.0224
Workout_Frequency	-2.396×10^{-5}	1.568×10^{-5}	-1.528	0.12655	-0.0111
Residual standard error: 0.001958 (df = 19995)					
Multiple R^2: 0.00223 *Adjusted* R^2: 0.00203					
F-statistic: 11.17 on 4 and 19995 df *p-value*: 4.74×10^{-9}					

Table 10 Generalized Additive Model (GAM) for Resting Heart Rate

Parametric Term	Estimate	Std. Error	t-value	p-value
Intercept	62.71797	0.22000	285.084	$< 2 \times 10^{-16}$ ***
Gender (Male)	0.05642	0.10207	0.553	0.58044
Workout_Frequency	-0.16581	0.06264	-2.647	0.00812 **
Approximate Significance of Smooth Terms				
Smooth Term	**edf**	**Ref.df**	**F**	**p-value**
s(Age)	7.439	8.400	11.94	$< 2 \times 10^{-16}$ ***
s(BMI)	10.392	10.902	16.87	$< 2 \times 10^{-16}$ ***
s(Water_Intake)	8.525	8.936	16.76	$< 2 \times 10^{-16}$ ***
Adjusted R^2: 0.0211 *Deviance Explained*: 2.25%				
REML: 67951 *Scale Estimate*: 52.012 n = 20,000				

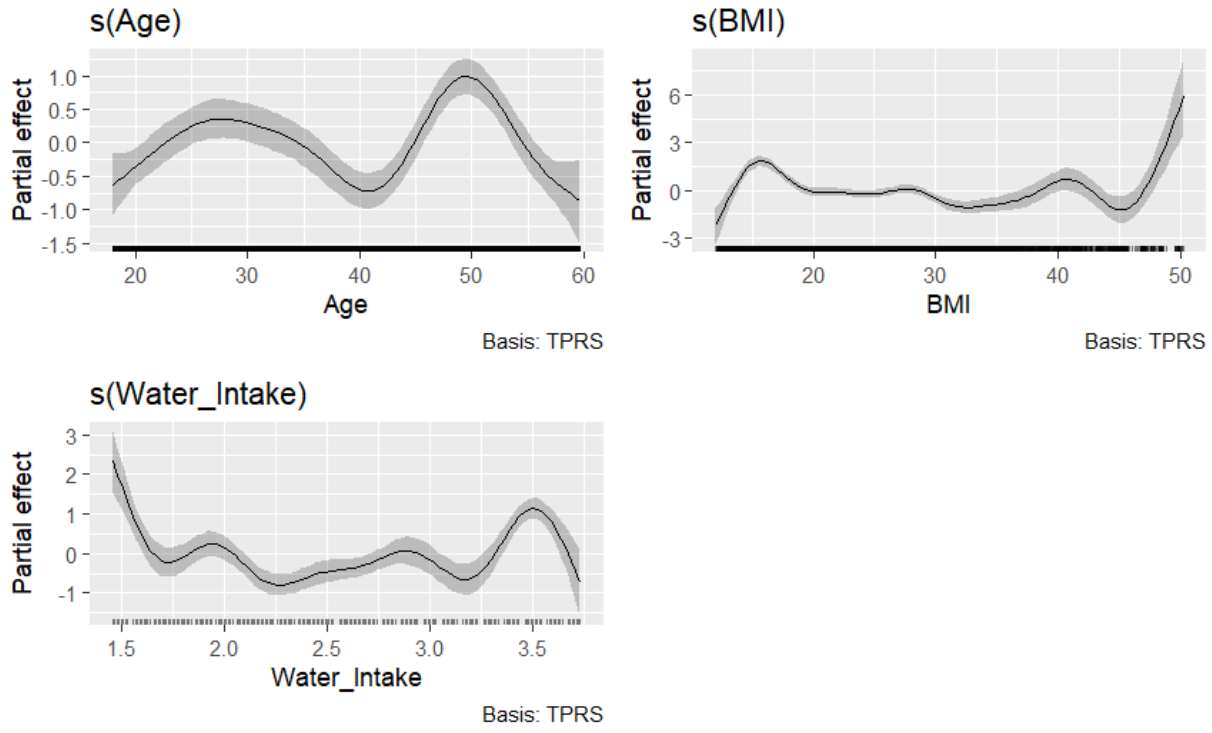


Fig. 10 Estimated smooth functions from the Generalized Additive Model (GAM). Panels show the partial effects of Age, BMI, and Water_Intake on Resting Heart Rate. Shaded regions indicate 95% confidence intervals. Basis type: Thin Plate Regression Splines (TPRS).

Table 11 Visual Inspection of Partial Effect Sizes from GAM Smooth Terms

In the original non-transformed linear model (Table 8), BMI and Water Intake were found to be statistically significant; however, the effect sizes were small, with Water Intake showing the largest effect (a 0.28 BPM increase in RHR per unit).

In the inverse linear model, BMI remained significant, while Water Intake was not statistically significant at the 5% level (two-sided). Although the coefficients of the inverse transformation are difficult to interpret directly, standardized estimates confirm that the effect sizes remain negligible.

In the Generalized Additive Model (GAM), no fixed-effect parameters were significant at the 5% level (two-sided). While the smoothed terms were statistically significant, the practical effect sizes were again small, as shown in Figure 10. The effects oscillated mostly within a range of ± 1 BPM, with spikes occurring only at the boundary values.

D. Post Model Analyses

1. Linear Models

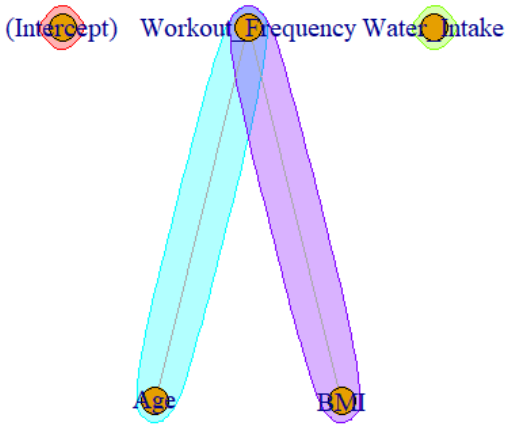


Fig. 11 Post-hoc comparison of linear model coefficients showing confidence ellipses and pairwise coefficient relationships.

The post hoc analysis of both the original and inverse linear models reveals that the effect sizes for Workout Frequency are comparable to those of Age and BMI. Furthermore, while the effects of Age, BMI, and Water Intake are statistically distinct from one another, they all remain small.

2. GAM

Table 12 GAM Basis Dimension (k) Checking Results

Smooth Term	k'	edf	k-index	p-value
s(Age)	9.00	7.44	0.97	0.02 *
s(BMI)	11.00	10.39	0.92	< 2e-16 ***
s(Water_Intake)	9.00	8.53	1.00	0.44

The K index table for the GAM fitted to the shows close values to the K-1 edf limit, even after tuning. However, if increased the smooths show wildly oscillatory behavior indicating the model is most likely chasing noise and edge cases rather than nonlinear trends in the data [7]. This indicates to me that despite significance being found in all the smooths the effect sizes remain practically small indicating no rigorously discernible influence on RHR from the smoothed predictors.

V. Discussion

A. Interpretation

The data and associated models developed for this analysis all point to one conclusion: the null result. Despite finding statistical significance in all models with very small p-values ($p < 0.025$), a deeper examination reveals that this significance is a testament to the large sample size rather than strong predictor influence (Table 8, Table 9, and Figure 10).

In addition to the small effect sizes, all models fitted to the data yielded residuals that did not follow a normal distribution, and both linear models displayed clear signs of heteroscedasticity. Although linear regression is robust to violations of normality, the original linear model contained influential points that can bias p-values and affect the rate of Type I errors. Furthermore, the variance explained by both models was $\approx 0.2\%$, highlighting the need for higher-quality data and/or more relevant predictors. Given these limitations, it is best to keep the null hypothesis [8].

The GAM detected strong effects near the minimum and maximum values of the RHR data (Figure 10). However, the predictor distribution plots (Figure 8) reveal that the largest effects coincided with the most skewed regions of the predictors, suggesting the model may be influenced by outliers in these boundary cases [8]. This region of the RHR data was also where the GAM deviated most strongly from normality (Figure 3), exhibiting the same heavy tails observed in the original RHR distribution. This casts doubt on the reliability of the GAM's results in these ranges [7]. As a result caution suggests siding with the null hypothesis, as it is difficult to defend that the model is capturing real trends rather than noise or outliers.

Ultimately, we failed to find statistically sound evidence that any single predictor in this dataset has a strong influence on RHR.

B. Practical Implications

Comparing these results with the literature reveals some concerning trends. We found no reporting of diagnostic plots or discussion regarding the limitations of linear model fits in similar studies. Furthermore, there was little discussion of the data distribution beyond summary statistics; even when reported, confidence intervals for means (e.g., for factors like Gender) often overlapped and showed no clear signs of being statistically different [9].

Another notable observation from the literature is the magnitude of reported effect sizes, which range around ≈ 1.5 BPM over a decade [10]. These effect sizes are similar to those observed in the GAM from this study. This raises the question of whether the introduction of large, readily available datasets from wearable technology contributes to an inflation of "statistical significance" reporting due to massive sample sizes, despite the absence of practical significance or a defined line for said practical significance.

C. Future Work

Given the limited causal inference inherent in cross-sectional studies—especially with data as highly variable as biometric measurements, analyzing longitudinal data with the same predictors could prove insightful. A repeated measures study fitted with a mixed-effects model might extract stronger signals and explain more variability. For instance, inter-individual heart rate can vary by as much as 70 BPM, yet remains much more stable within each individual over time [2]. In addition, finding a dataset including smoking habits, drinking habits, sleep, if on medication, and stress would be highly valuable, as these factors have been shown to influence RHR

One transformation that could improve linear model fits is the Rank-Based Inverse Normal Transformation [8]. Future work could involve performing this transformation to account for heavy tails, followed by fitting another linear model to determine if it explains more variance, yields stronger effect sizes, and improves diagnostic plots. However, other studies have found nonlinearities in this type of data, which might limit the improvement offered by linear techniques [2].

Finally, since we found statistical significance but retained the null hypothesis due to small effect sizes (a pattern also seen in the literature), performing a sensitivity or power analysis is essential. This could be conducted utilizing the `pwr` package in R. Quantifying the power achieved in this analysis would help clarify the risk of Type I and Type II errors. This step could lead to a reevaluation of the conclusions drawn from this dataset and validate the reliability of the findings.

VI. Conclusion

Utilizing a large dataset from Kaggle ($n = 20,000$), we performed a statistical analysis of RHR data and found it to be not normally distributed. The RHR data exhibited heavy tails reminiscent of a Cauchy or low degrees-of-freedom (DOF) Student's t-distribution

The distributions of RHR across gender and weekly workout frequency groups followed the same pattern as the overall RHR distribution; consequently, non-parametric statistical tests were performed, alongside parametric tests to assess robustness.

After a correlation analysis reduced the set of predictors, the distributions of the remaining variables were plotted, displaying high levels of skewness and kurtosis. Subsequently, two linear models and one Generalized Additive Model (GAM) were fitted to the data, all of which demonstrated poor performance. All models explained a negligible amount of the variance in the data (maximum 2%), and both linear models exhibited poor diagnostic plots (non-normal residuals, heteroscedasticity, and influential points).

Comparison with the literature revealed similar results; however, validating these findings was difficult due to the limited statistical reporting and lack of diagnostic transparency in RHR studies.

Appendix

Project Code

Angel Rodas

2025-11-06

Data Cleaning and Curating

```
# Loading the Data
lifestyle_data <- read_csv("C:\\Users\\ADRod\\Downloads\\archive\\Final_data.csv",
                           show_col_types = FALSE)

# Rename
clean <- lifestyle_data %>%
  rename(
    Workout_Frequency = `Workout_Frequency (days/week)`,
    Water_Intake      = `Water_Intake (liters)`
  ) %>%
  dplyr::select(
    Resting_BPM, Age, Gender, BMI, Fat_Percentage,
    Workout_Frequency, Calories, Water_Intake
  ) %>%
  mutate(Gender = factor(Gender)) %>%
  mutate(Workout_Frequency = round(Workout_Frequency)) %>%
  mutate(Workout_Frequency = as.numeric(Workout_Frequency))

my_data <- clean

summary(my_data)
```

```
##   Resting_BPM         Age           Gender           BMI        Fat_Percentage
##  Min.   :49.49   Min.   :18.00   Female:10028   Min.   :12.04   Min.   :11.33
##  1st Qu.:55.96   1st Qu.:28.17   Male  : 9972   1st Qu.:20.10   1st Qu.:22.39
##  Median :62.20   Median :39.87                  Median :24.12   Median :25.82
##  Mean   :62.20   Mean   :38.85                  Mean   :24.92   Mean   :26.10
##  3rd Qu.:68.09   3rd Qu.:49.63                  3rd Qu.:28.56   3rd Qu.:29.68
##  Max.   :74.50   Max.   :59.67                  Max.   :50.23   Max.   :35.00
##  Workout_Frequency    Calories     Water_Intake
##  Min.   :2.000     Min.   : 781   Min.   :1.460
##  1st Qu.:3.000     1st Qu.:1634   1st Qu.:2.170
##  Median :3.000     Median :1919   Median :2.610
##  Mean   :3.319     Mean   :2024   Mean   :2.627
##  3rd Qu.:4.000     3rd Qu.:2360   3rd Qu.:3.120
##  Max.   :5.000     Max.   :3641   Max.   :3.730
```

```r
str(my_data)
```

```
## tibble [20,000 x 8] (S3: tbl_df/tbl/data.frame)
##  $ Resting_BPM      : num [1:20000] 69 73.2 55 50.1 70.8 ...
##  $ Age              : num [1:20000] 34.9 23.4 33.2 38.7 45.1 ...
##  $ Gender           : Factor w/ 2 levels "Female","Male": 2 1 1 1 2 1 2 1 1 1 ...
##  $ BMI              : num [1:20000] 24.9 23.5 21.1 32.5 14.8 ...
##  $ Fat_Percentage   : num [1:20000] 26.8 27.7 24.3 32.8 17.3 ...
##  $ Workout_Frequency: num [1:20000] 4 4 3 4 4 3 5 4 4 2 ...
##  $ Calories         : num [1:20000] 1806 1577 1608 2657 1470 ...
##  $ Water_Intake     : num [1:20000] 1.5 1.9 1.88 2.5 2.91 2.91 2.71 2.88 3.49 2.49 ...
```

```r
# Missingness
colSums(is.na(my_data))
```

```
##       Resting_BPM               Age            Gender               BMI
##                 0                 0                 0                 0
##    Fat_Percentage Workout_Frequency          Calories      Water_Intake
##                 0                 0                 0                 0
```

```r
my_data <- na.omit(my_data)

# Remove Outliers
# <- my_data %>%
  #filter(Resting_BPM >= 40, Resting_BPM <= 120,
         #BMI >= 12, BMI <= 50)

# Quick summary
summary(my_data)
```

```
##   Resting_BPM         Age           Gender           BMI        Fat_Percentage
##  Min.   :49.49   Min.   :18.00   Female:10028   Min.   :12.04   Min.   :11.33
##  1st Qu.:55.96   1st Qu.:28.17   Male  : 9972   1st Qu.:20.10   1st Qu.:22.39
##  Median :62.20   Median :39.87                  Median :24.12   Median :25.82
##  Mean   :62.20   Mean   :38.85                  Mean   :24.92   Mean   :26.10
##  3rd Qu.:68.09   3rd Qu.:49.63                  3rd Qu.:28.56   3rd Qu.:29.68
##  Max.   :74.50   Max.   :59.67                  Max.   :50.23   Max.   :35.00
##  Workout_Frequency    Calories     Water_Intake
##  Min.   :2.000     Min.   : 781   Min.   :1.460
##  1st Qu.:3.000     1st Qu.:1634   1st Qu.:2.170
##  Median :3.000     Median :1919   Median :2.610
##  Mean   :3.319     Mean   :2024   Mean   :2.627
##  3rd Qu.:4.000     3rd Qu.:2360   3rd Qu.:3.120
##  Max.   :5.000     Max.   :3641   Max.   :3.730
```

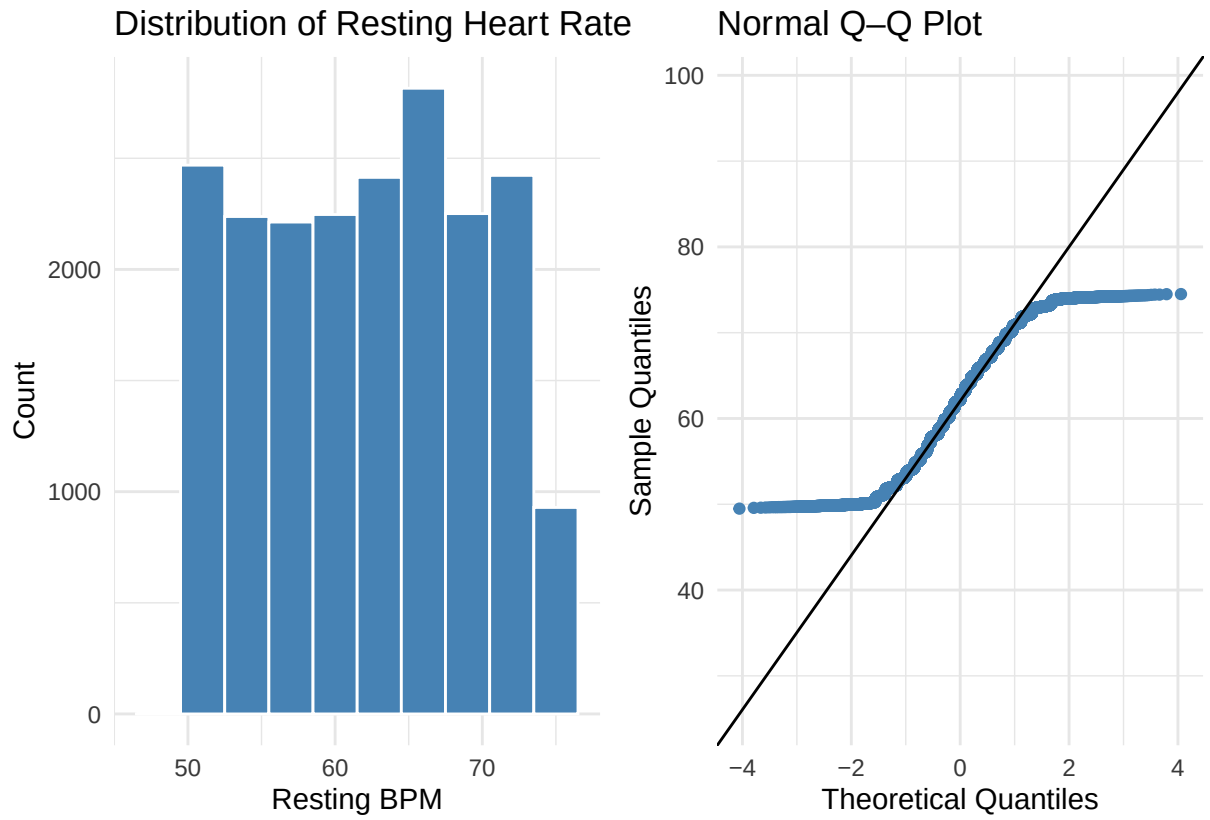
```r
str(my_data)
```

```
## tibble [20,000 x 8] (S3: tbl_df/tbl/data.frame)
##  $ Resting_BPM      : num [1:20000] 69 73.2 55 50.1 70.8 ...
##  $ Age              : num [1:20000] 34.9 23.4 33.2 38.7 45.1 ...
##  $ Gender           : Factor w/ 2 levels "Female","Male": 2 1 1 1 2 1 2 1 1 1 ...
```

```
##  $ BMI              : num [1:20000] 24.9 23.5 21.1 32.5 14.8 ...
##  $ Fat_Percentage   : num [1:20000] 26.8 27.7 24.3 32.8 17.3 ...
##  $ Workout_Frequency: num [1:20000] 4 4 3 4 4 3 5 4 4 2 ...
##  $ Calories         : num [1:20000] 1806 1577 1608 2657 1470 ...
##  $ Water_Intake     : num [1:20000] 1.5 1.9 1.88 2.5 2.91 2.91 2.71 2.88 3.49 2.49 ...
```

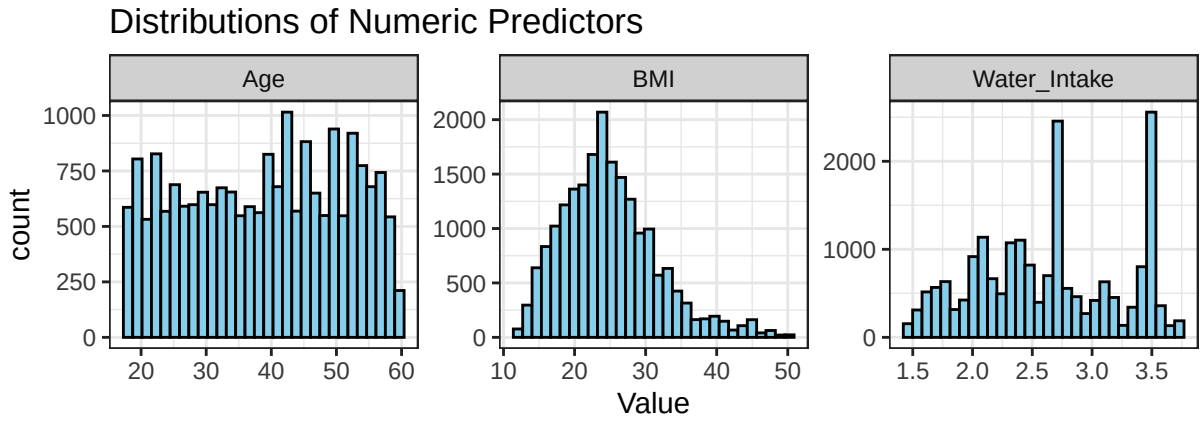
Visualization

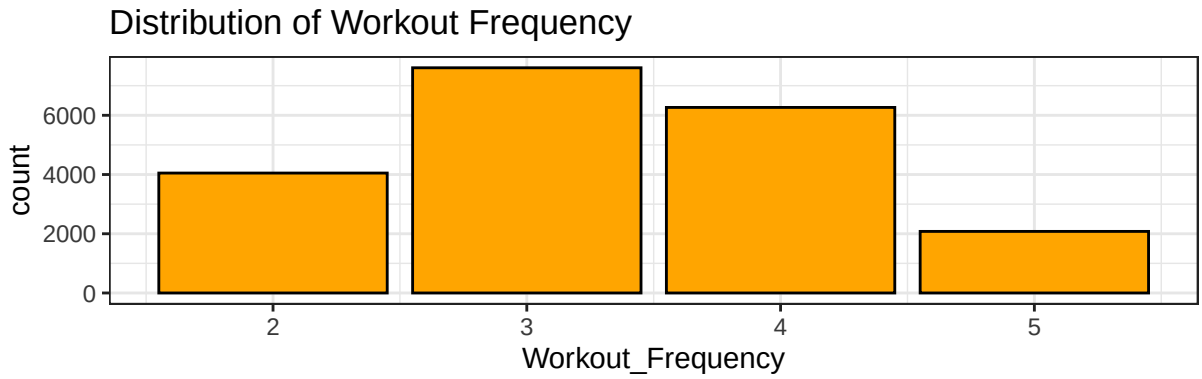
```
# Histogram (your original plot)
p1 <- ggplot(my_data, aes(x = Resting_BPM)) +
  geom_histogram(binwidth = 3, fill = "steelblue", color = "white") +
  labs(title = "Distribution of Resting Heart Rate",
       x = "Resting BPM", y = "Count") +
  theme_minimal()

# QQ Plot using ggplot2 instead of base R
p2 <- ggplot(my_data, aes(sample = Resting_BPM)) +
  stat_qq(color = "steelblue") +
  stat_qq_line(color = "black") +
  labs(title = "Normal Q-Q Plot",
       x = "Theoretical Quantiles", y = "Sample Quantiles") +
  theme_minimal()

# Combine using patchwork
p1 | p2
```



Dataset too Large to do shapiro wilk, however QQ plot shows clear deviations from normality.

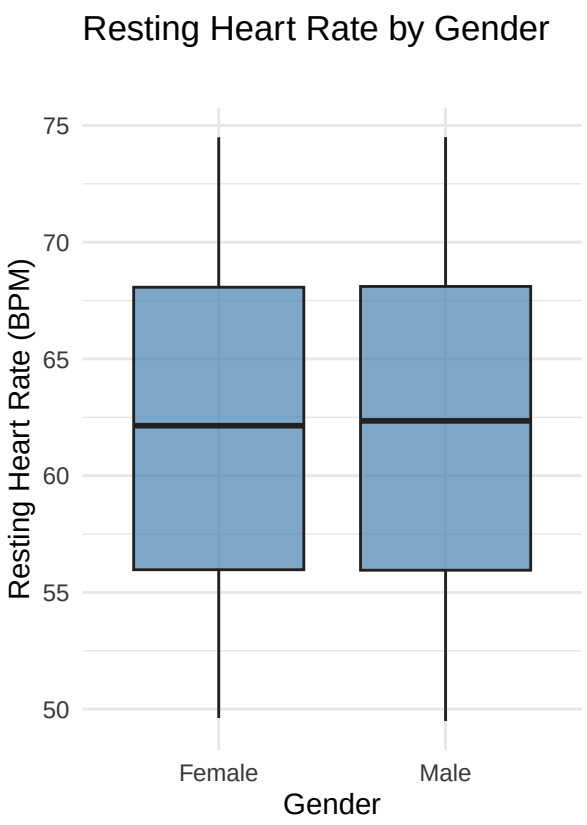
```
# Select predictors from your model
predictors <- dplyr::select(my_data, Age, BMI, Workout_Frequency, Water_Intake)


# ---- Histogram for numeric variables ----
p_num <- predictors %>%
  dplyr::select(-Workout_Frequency) %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Value") %>%
  ggplot(aes(x = Value)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black") +
  facet_wrap(~ Variable, scales = "free") +
  theme_bw() +
  labs(title = "Distributions of Numeric Predictors")

# ---- Bar plot for Workout_Frequency (factor) ----
p_factor <- predictors %>%
  ggplot(aes(x = Workout_Frequency)) +
  geom_bar(fill = "orange", color = "black") +
  theme_bw() +
  labs(title = "Distribution of Workout Frequency")

# Display combined plot (numeric + factor)
p_num / p_factor
```

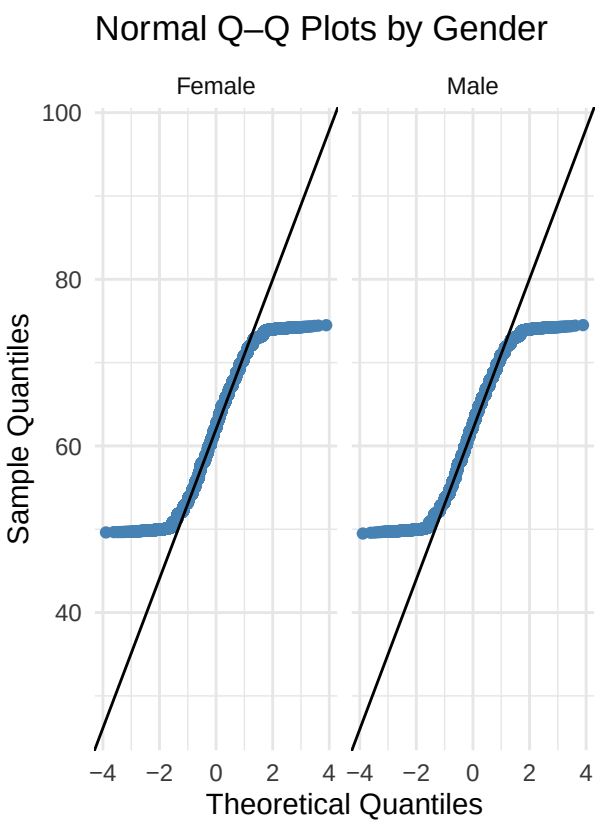




Gender effect on RHR side tnagent

```
p1 <- ggplot(my_data, aes(x = Gender, y = Resting_BPM)) +
  geom_boxplot(fill = "steelblue", alpha = 0.7) +
  labs(
    title = "Resting Heart Rate by Gender",
    x = "Gender",
    y = "Resting Heart Rate (BPM)"
  ) +
  theme_minimal()

p2 <- ggplot(my_data, aes(sample = Resting_BPM)) +
  stat_qq(color = "steelblue") +
  stat_qq_line(color = "black") +
  facet_wrap(~ Gender) +
  labs(
    title = "Normal Q-Q Plots by Gender",
    x = "Theoretical Quantiles",
    y = "Sample Quantiles"
  ) +
  theme_minimal()

p1 | p2
```





Non-parametric test to see if Male and Female RHR is the same

```
summary(my_data$Gender)
```

```
## Female   Male
##  10028   9972
```

```
gender_anova_model <- aov(Resting_BPM ~ Gender, data = my_data)
summary(gender_anova_model)
```

```
##                Df  Sum Sq Mean Sq F value Pr(>F)
## Gender          1       8    8.46   0.159   0.69
## Residuals   19998 1062631   53.14
```

```
wilcox.test(Resting_BPM ~ Gender, data = my_data)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Resting_BPM by Gender
## W = 49841984, p-value = 0.6994
## alternative hypothesis: true location shift is not equal to 0
```

```r
leveneTest(Resting_BPM ~ Gender, data = my_data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value Pr(>F)
## group     1  0.3823 0.5364
##       19998
```

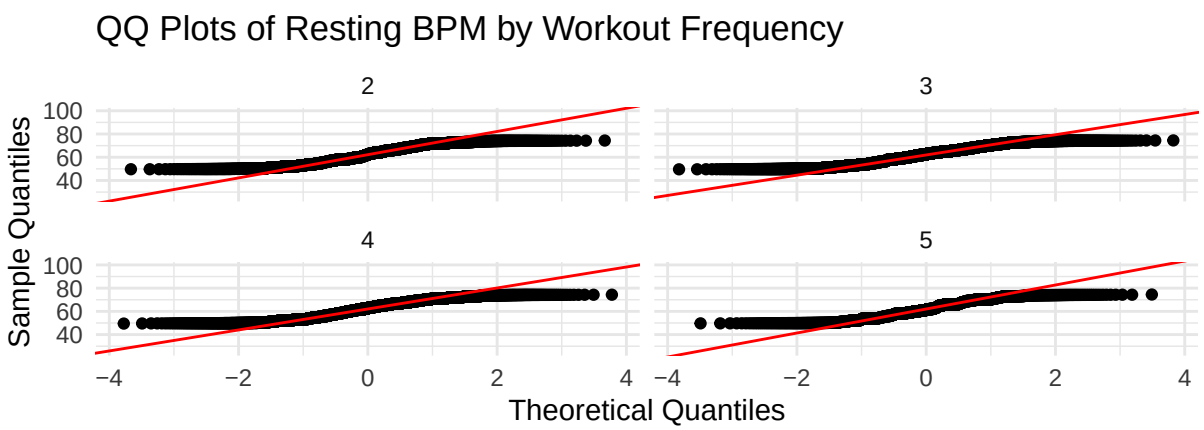
Variance is homogenous.

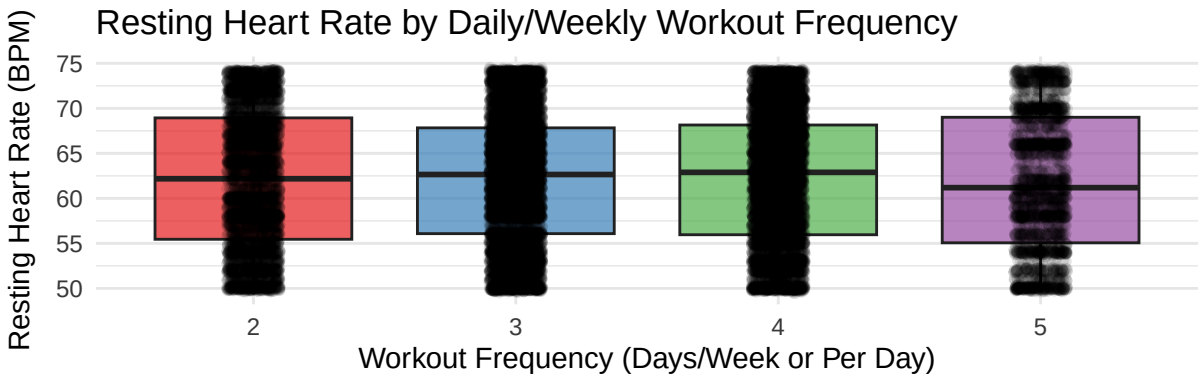
Comparing workout frequency as a Factor

```r
my_data_filtered <- my_data %>%
  filter(Workout_Frequency %in% c(2, 3, 4, 5))

my_data_filtered$Workout_Frequency <- factor(my_data_filtered$Workout_Frequency)

summary(my_data_filtered$Workout_Frequency)
```

```
##    2    3    4    5
## 4050 7605 6266 2079
```

```r
# Create QQ plot
p_qq <- ggplot(my_data_filtered, aes(sample = Resting_BPM)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  facet_wrap(~ Workout_Frequency) +
  labs(
    title = "QQ Plots of Resting BPM by Workout Frequency",
    x = "Theoretical Quantiles",
    y = "Sample Quantiles"
  ) +
  theme_minimal()

# Create box+jitter plot
p_bpm_by_workout <- ggplot(my_data_filtered, aes(x = Workout_Frequency, y = Resting_BPM, fill = Workout
  geom_boxplot(alpha = 0.7, outlier.shape = 1) +
  geom_jitter(width = 0.1, alpha = 0.1, color = "black") +
  labs(
    title = "Resting Heart Rate by Daily/Weekly Workout Frequency",
    x = "Workout Frequency (Days/Week or Per Day)",
    y = "Resting Heart Rate (BPM)"
  ) +
  scale_fill_brewer(palette = "Set1") +
  theme_minimal() +
  theme(legend.position = "none")

# Combine the two plots vertically
p_qq / p_bpm_by_workout
```






```r
workout_frequency_anova_model <- aov(Resting_BPM ~ Workout_Frequency, data = my_data_filtered)

summary(workout_frequency_anova_model)
```

```
##                      Df  Sum Sq Mean Sq F value Pr(>F)
## Workout_Frequency     3     222   73.94   1.392  0.243
## Residuals         19996 1062418   53.13
```

```r
leveneTest(Resting_BPM ~ Gender, data = my_data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value Pr(>F)
## group     1  0.3823 0.5364
##       19998
```

```r
tukey_hsd <- TukeyHSD(workout_frequency_anova_model)
print(tukey_hsd)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Resting_BPM ~ Workout_Frequency, data = my_data_filtered)
##
## $Workout_Frequency
```

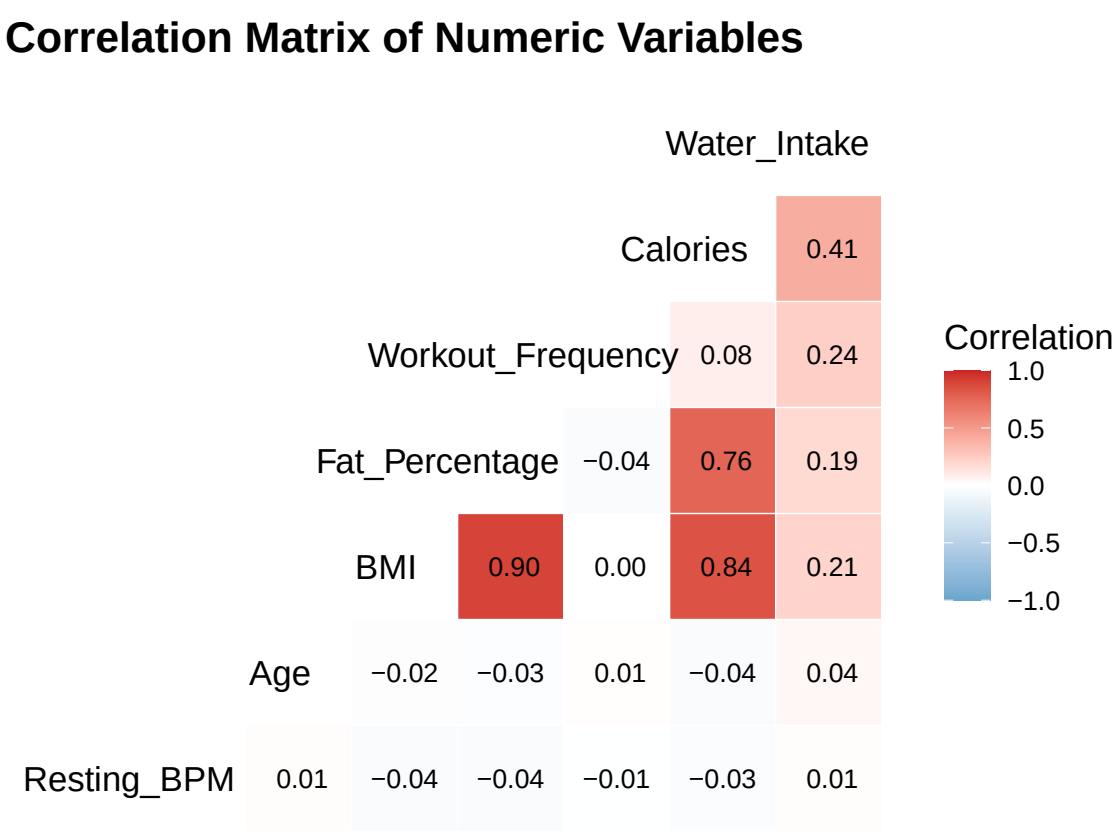
```
##            diff        lwr       upr     p adj
## 3-2  0.03785394 -0.3264477 0.4021556 0.9933455
## 4-2  0.08660094 -0.2909843 0.4641862 0.9353611
## 5-2 -0.28329809 -0.7885665 0.2219703 0.4739529
## 4-3  0.04874699 -0.2707679 0.3682619 0.9795714
## 5-3 -0.32115203 -0.7846338 0.1423297 0.2828492
## 5-4 -0.36989903 -0.8438930 0.1040949 0.1860785
```

```r
kruskal.test(Resting_BPM ~ Workout_Frequency, data = my_data_filtered)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Resting_BPM by Workout_Frequency
## Kruskal-Wallis chi-squared = 3.8968, df = 3, p-value = 0.2728
```

Checking Predictor Correlation

```r
my_data_num <- my_data %>% dplyr::select(-Gender)

GGally::ggcorr(
  my_data_num,
  label = TRUE,
  label_round = 2,
  label_size = 3.5,         # Larger, clearer labels
  hjust = 0.8,              # Adjust horizontal alignment
  layout.exp = 2,           # Expands the plot area for spacing
  low = "skyblue3",         # Better low color
  mid = "white",
  high = "firebrick3",      # Better high color
  name = "Correlation"      # Legend title
) +
  ggtitle("Correlation Matrix of Numeric Variables") +
  theme_minimal(base_size = 13) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 11),
    axis.text.y = element_text(size = 11),
    plot.title = element_text(hjust = 0.5, face = "bold")
  )
```



Removing fat percentage and calorie intake from predictors.

```
my_data <- my_data %>%
  dplyr::select(Resting_BPM, Age, Gender, BMI,
                Workout_Frequency, Water_Intake)
```

Plotting the remaining variables

```
p1 <- ggplot(my_data, aes(x = Age, y = Resting_BPM, color = Gender)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(title = "Resting Heart Rate vs Age", x = "Age", y = "Resting BPM") +
  theme_minimal()

p2 <- ggplot(my_data, aes(x = BMI, y = Resting_BPM, color = Gender)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(title = "Resting Heart Rate vs BMI", x = "BMI", y = "Resting BPM") +
  theme_minimal()

p3 <- ggplot(my_data, aes(x = Workout_Frequency, y = Resting_BPM, color = Gender)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(title = "Resting Heart Rate vs Workout Frequency",
       x = "Workout Frequency (days/week)", y = "Resting BPM") +
```

```
  theme_minimal()

p4 <- ggplot(my_data, aes(x = Water_Intake, y = Resting_BPM, color = Gender)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(title = "Resting Heart Rate vs Water Intake",
       x = "Water Intake (liters)", y = "Resting BPM") +
  theme_minimal()
(p1 | p2) /
(p3 | p4)
```

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

Setting up 10-cross fold validation

```
# Set seed for reproducibility
set.seed(123)
# Set up repeated k-fold cross-validation
train.control <- trainControl(method = "cv", number = 10)
```

```
# Train the model
step.model <- train(Resting_BPM ~., data = my_data,
                    method = "leapSeq",
                    tuneGrid = data.frame(nvmax = 1:6),
                    trControl = train.control,
                    na.action=na.exclude
)
step.model$results
```

```
##   nvmax     RMSE    Rsquared      MAE     RMSESD  RsquaredSD      MAESD
## 1     1 7.283415 0.002070062 6.279853 0.03967434 0.001548463 0.04846262
## 2     2 7.282043 0.002520297 6.279377 0.04021316 0.002199775 0.04916722
## 3     3 7.284303 0.001817760 6.280237 0.03985374 0.001470892 0.04846401
## 4     4 7.281641 0.002705359 6.277306 0.04071225 0.002337141 0.04991892
## 5     5 7.281950 0.002567575 6.277956 0.04048457 0.002236031 0.04929168
## 6     6 7.281950 0.002567575 6.277956 0.04048457 0.002236031 0.04929168
```

Finding Optimal Model Parameters

```
step.model$bestTune
```

```
##   nvmax
## 4     4
```

Finding which predictors worked the best

```
summary(step.model$finalModel)
```

```
## Subset selection object
## 5 Variables  (and intercept)
##                   Forced in Forced out
## Age                   FALSE      FALSE
## GenderMale            FALSE      FALSE
## BMI                   FALSE      FALSE
## Workout_Frequency     FALSE      FALSE
## Water_Intake          FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: 'sequential replacement'
##          Age GenderMale BMI Workout_Frequency Water_Intake
## 1  ( 1 ) " " " "        "*" " "               " "
## 2  ( 1 ) " " " "        "*" " "               "*"
## 3  ( 1 ) "*" "*"        "*" " "               " "
## 4  ( 1 ) "*" " "        "*" "*"               "*"
```

Fitting the four parameter linear model

```
m.final <- lm(Resting_BPM ~ Age + BMI + Workout_Frequency + Water_Intake, data = my_data)
summary(m.final)
```

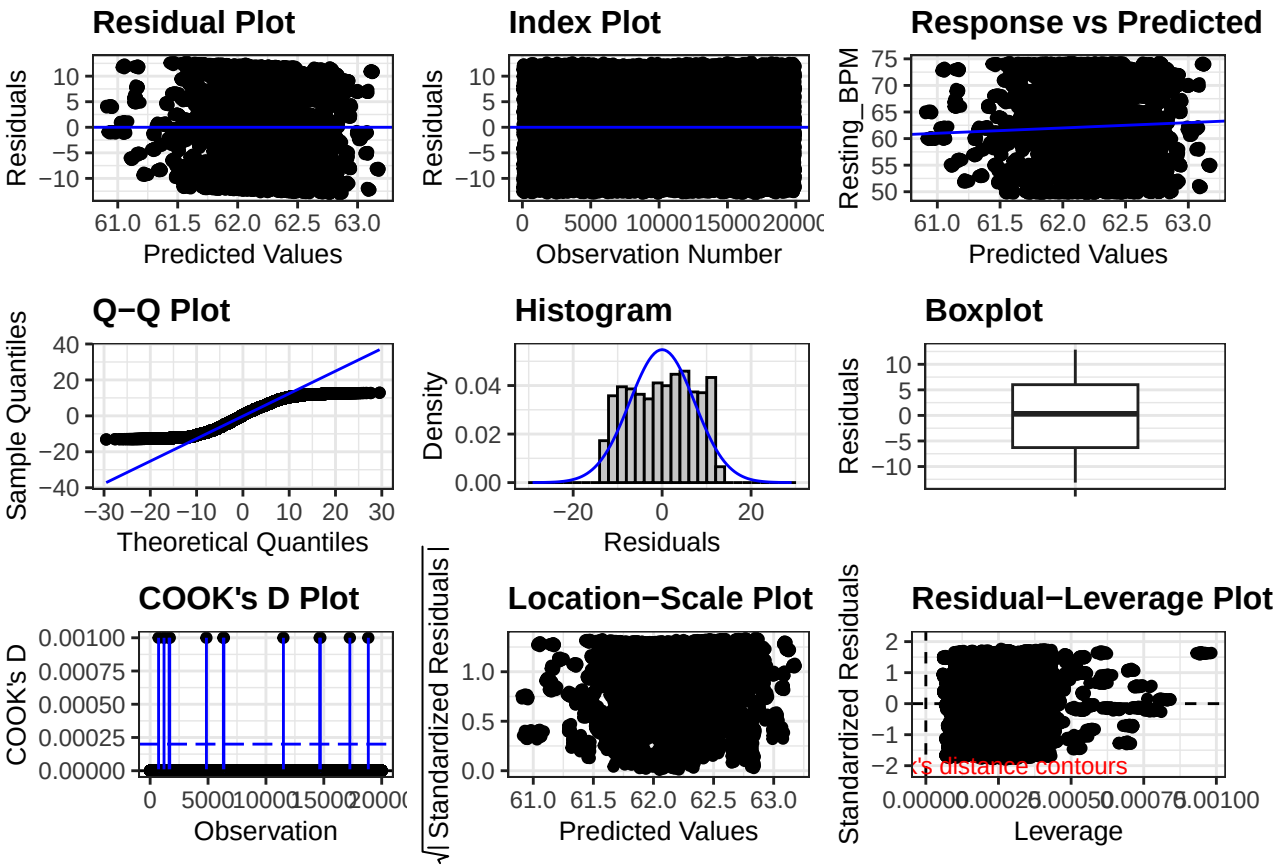
```
##
## Call:
## lm(formula = Resting_BPM ~ Age + BMI + Workout_Frequency + Water_Intake,
##     data = my_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.1393  -6.2960   0.3075   6.0118  12.8646
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       62.701654   0.348603 179.866  < 2e-16 ***
## Age                0.007103   0.004255   1.670  0.09502 .
## BMI               -0.049601   0.007880  -6.294 3.15e-10 ***
## Workout_Frequency -0.089121   0.058321  -1.528  0.12650
## Water_Intake       0.285476   0.090030   3.171  0.00152 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.282 on 19995 degrees of freedom
## Multiple R-squared:  0.002352,   Adjusted R-squared:  0.002152
## F-statistic: 11.78 on 4 and 19995 DF,  p-value: 1.467e-09
```

Diagnostic Plots

```
resid_panel(m.final, plots = "all")
```

```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## i The deprecated feature was likely used in the ggResidpanel package.
##   Please report the issue to the authors.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## i The deprecated feature was likely used in the ggResidpanel package.
##   Please report the issue to the authors.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



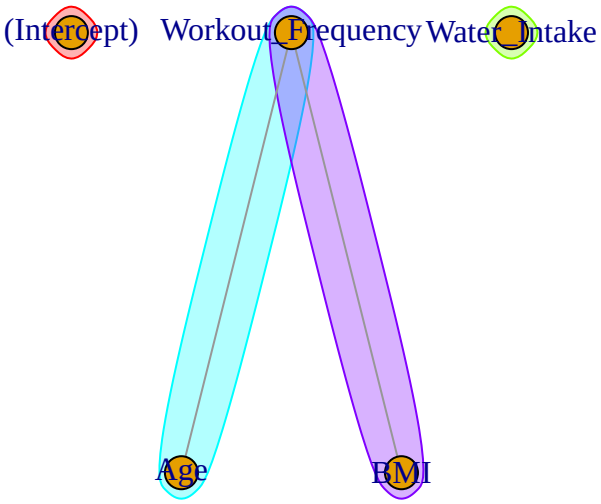
Post Hoc Analysis

```
ZZ <- posthoc(Model = m.final, digits = 3, padjust = "holm")
summary(ZZ)
```

```
##                      Est.   2.5%  97.5% Group (Intercept)   Age   BMI
## (Intercept)       62.702 62.018 63.385     d
## Age                0.007 -0.001  0.015     b           0
## BMI               -0.050 -0.065 -0.034     a           0     0
## Workout_Frequency -0.089 -0.203  0.025    ab           0 0.354  0.93
## Water_Intake       0.285  0.109  0.462     c           0  0.01 0.001
##                   Workout_Frequency
## (Intercept)
## Age
## BMI
## Workout_Frequency
## Water_Intake                  0.008
```

Comparing the different factor levels

```
plot(ZZ)
```



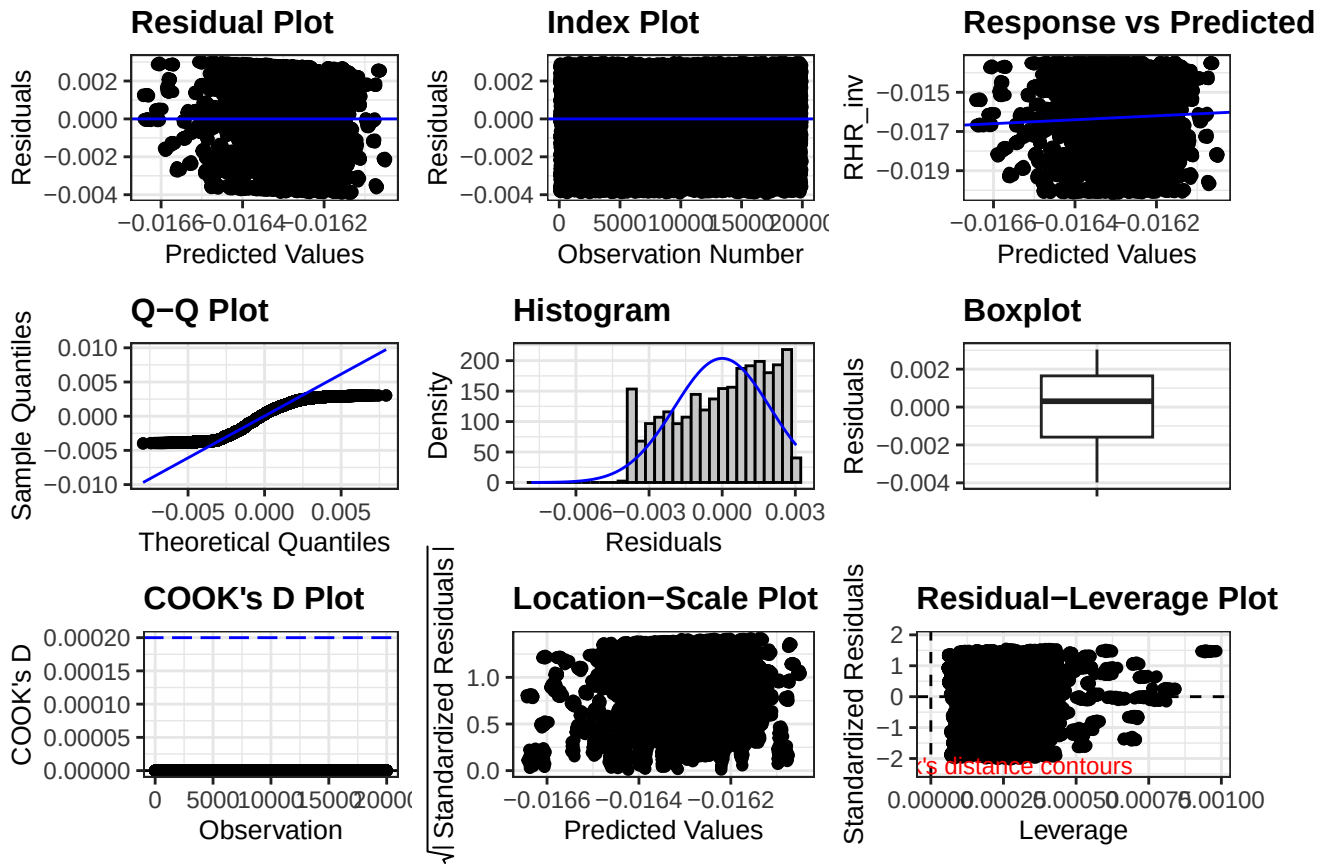
Applying the negative inverse transform

```
# Negative inverse transform
my_data$RHR_inv <- - 1 / my_data$Resting_BPM

# Fit model on transformed outcome
m.final.inv <- lm(
  RHR_inv ~ Age + BMI + Water_Intake + Workout_Frequency,
  data = my_data
)

summary(m.final.inv)
```

```
##
## Call:
## lm(formula = RHR_inv ~ Age + BMI + Water_Intake + Workout_Frequency,
##     data = my_data)
```

```
##
## Residuals:
##        Min          1Q      Median          3Q         Max
## -0.0039800 -0.0015867  0.0003071  0.0016462  0.0030360
##
## Coefficients:
##                      Estimate Std. Error  t value Pr(>|t|)
## (Intercept)       -1.617e-02  9.373e-05 -172.522  < 2e-16 ***
## Age                1.960e-06  1.144e-06    1.713  0.08672 .
## BMI               -1.297e-05  2.119e-06   -6.121 9.49e-10 ***
## Water_Intake       7.249e-05  2.421e-05    2.995  0.00275 **
## Workout_Frequency -2.396e-05  1.568e-05   -1.528  0.12655
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001958 on 19995 degrees of freedom
## Multiple R-squared:  0.00223,    Adjusted R-squared:  0.00203
## F-statistic: 11.17 on 4 and 19995 DF,  p-value: 4.74e-09
```

```
resid_panel(m.final.inv, plots = "all")
```




```
lm.beta(m.final.inv)
```

```
##              Age              BMI      Water_Intake Workout_Frequency
##       0.01211374      -0.04434224       0.02236906      -0.01113836
```

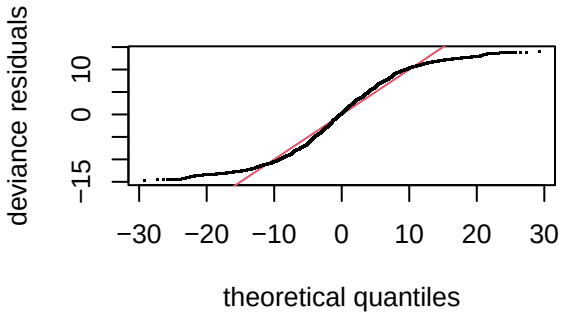
Applying a GAM for fitting the data

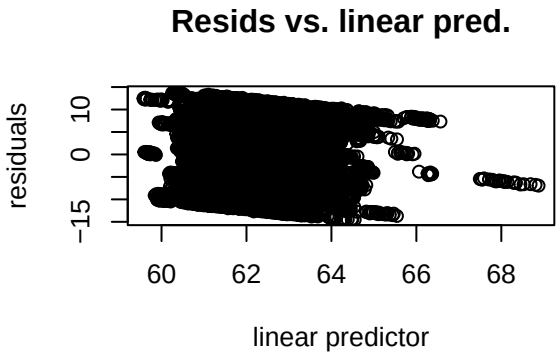
```
gam_fit <- gam(
  Resting_BPM ~
    s(Age, k = 10) +
    s(BMI, k = 12) +
    s(Water_Intake, k = 10) +
    Gender+
    Workout_Frequency,
  data = my_data,
  method = "REML"
)
```

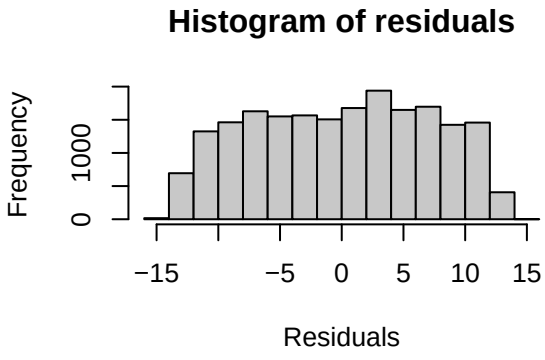
```
summary(gam_fit)
```

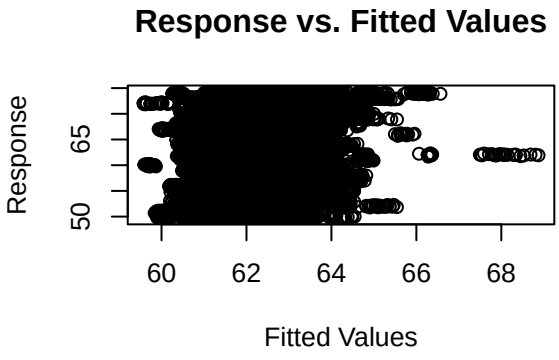
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Resting_BPM ~ s(Age, k = 10) + s(BMI, k = 12) + s(Water_Intake,
##     k = 10) + Gender + Workout_Frequency
##
## Parametric coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       62.71797    0.22000 285.084  < 2e-16 ***
## GenderMale         0.05642    0.10207   0.553  0.58044
## Workout_Frequency -0.16581    0.06264  -2.647  0.00812 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                     edf Ref.df     F p-value
## s(Age)            7.439  8.400 11.94  <2e-16 ***
## s(BMI)           10.392 10.902 16.87  <2e-16 ***
## s(Water_Intake)   8.525  8.936 16.76  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0211   Deviance explained = 2.25%
## -REML =  67951  Scale est. = 52.012    n = 20000
```

```
# Basic diagnostic checks
gam.check(gam_fit)
```



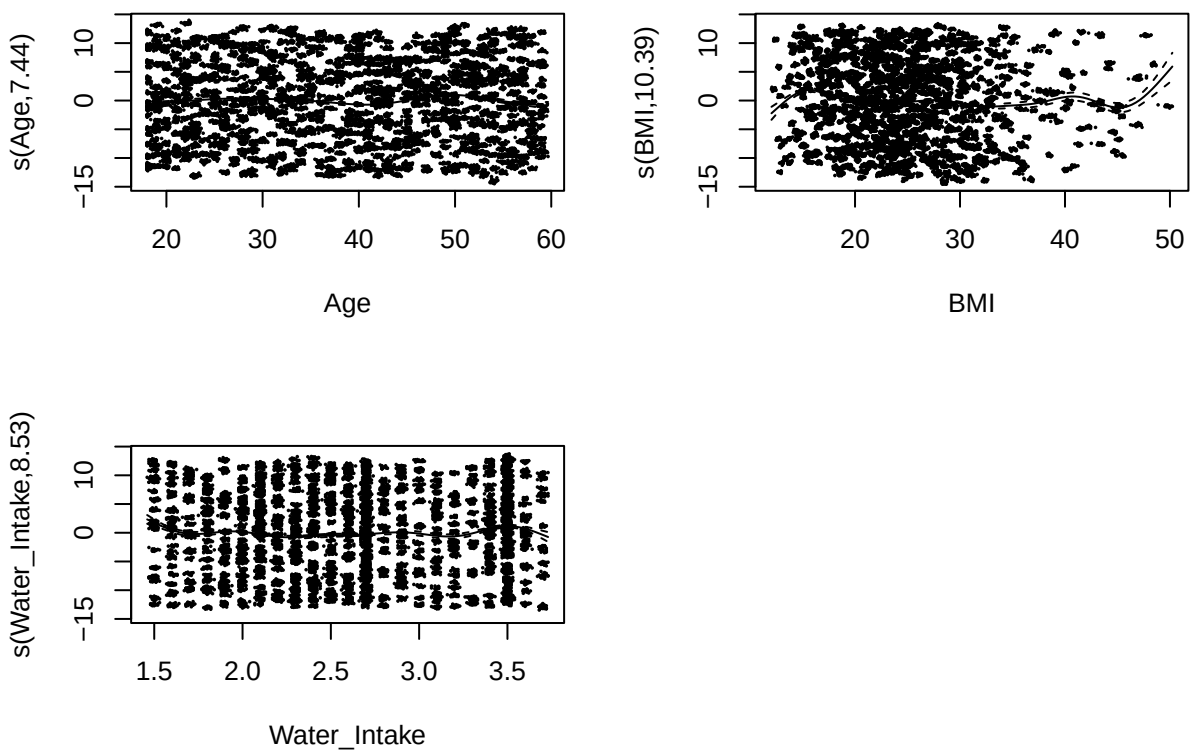







```
##
## Method: REML   Optimizer: outer newton
## full convergence after 8 iterations.
## Gradient range [-0.003462493,0.001477323]
## (score 67950.57 & scale 52.01209).
## Hessian positive definite, eigenvalue range [2.559479,9997.008].
## Model rank =  32 / 32
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##                   k'   edf k-index p-value
## s(Age)          9.00  7.44    0.97    0.02 *
## s(BMI)         11.00 10.39    0.92  <2e-16 ***
## s(Water_Intake)  9.00  8.53    1.00    0.44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

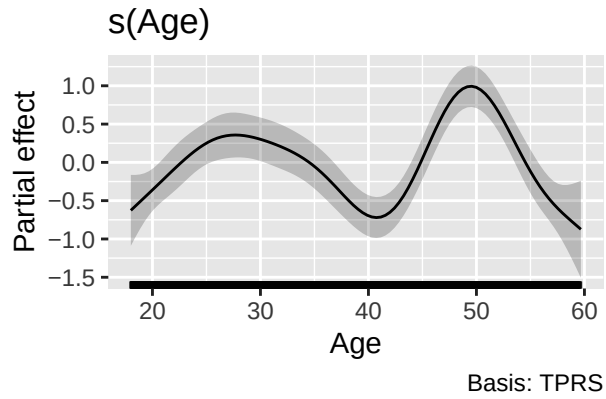
```r
# Plot smooths
par(mfrow = c(2, 3))
plot(gam_fit,
     residuals = TRUE,
     pch       = 16,
     cex       = 0.3,
     shade     = TRUE,
     pages     = 1)
```

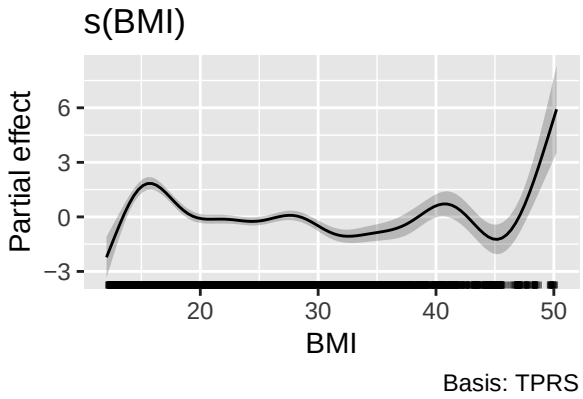


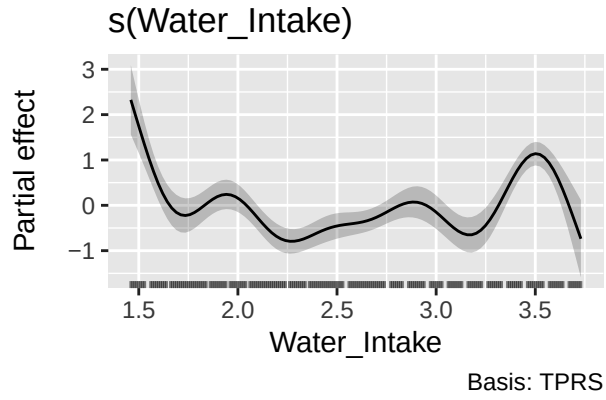

```
par(mfrow = c(1, 1))
```

Beter Plotting

```
draw(gam_fit,
     residuals = FALSE,
     scales    = "free")
```







Model Comparison

```
# AIC / BIC comparison
AIC(m.final, gam_fit)
```

```
##               df      AIC
## m.final  6.00000 136178.0
## gam_fit 31.43349 135820.6
```

```
BIC(m.final, gam_fit)
```

```
##               df      BIC
## m.final  6.00000 136225.5
## gam_fit 31.43349 136069.0
```

```
# Optional: approximate test between lm and GAM
anova(m.final, gam_fit, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: Resting_BPM ~ Age + BMI + Workout_Frequency + Water_Intake
```

```
## Model 2: Resting_BPM ~ s(Age, k = 10) + s(BMI, k = 12) + s(Water_Intake,
##     k = 10) + Gender + Workout_Frequency
##   Res.Df     RSS     Df Sum of Sq      F    Pr(>F)
## 1  19995 1060141
## 2  19971 1038715 24.356     21426 16.913 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# True outcome
y <- my_data$Resting_BPM

# Predictions on original BPM scale
pred_lm   <- predict(m.final)                 # already in BPM
pred_gam  <- predict(gam_fit)                 # already in BPM
pred_inv  <- predict(m.final.inv)             # on inverse scale
pred_inv_back <- -1 / pred_inv                # back-transform to BPM

# Helper to compute metrics
model_metrics <- function(obs, pred) {
  data.frame(
    RMSE = sqrt(mean((obs - pred)^2)),
    MAE  = mean(abs(obs - pred)),
    R2   = 1 - sum((obs - pred)^2) / sum((obs - mean(obs))^2)
  )
}

results <- bind_rows(
  Linear          = model_metrics(y, pred_lm),
  InverseLinear   = model_metrics(y, pred_inv_back),
  GAM             = model_metrics(y, pred_gam),
  .id = "Model"
)

results
```

```
##           Model     RMSE      MAE           R2
## 1        Linear 7.280593 6.276755  0.002351715
## 2 InverseLinear 7.332215 6.331437 -0.011845866
## 3           GAM 7.206646 6.199122  0.022514316
```

```
BIC(m.final, m.final.inv , gam_fit)
```

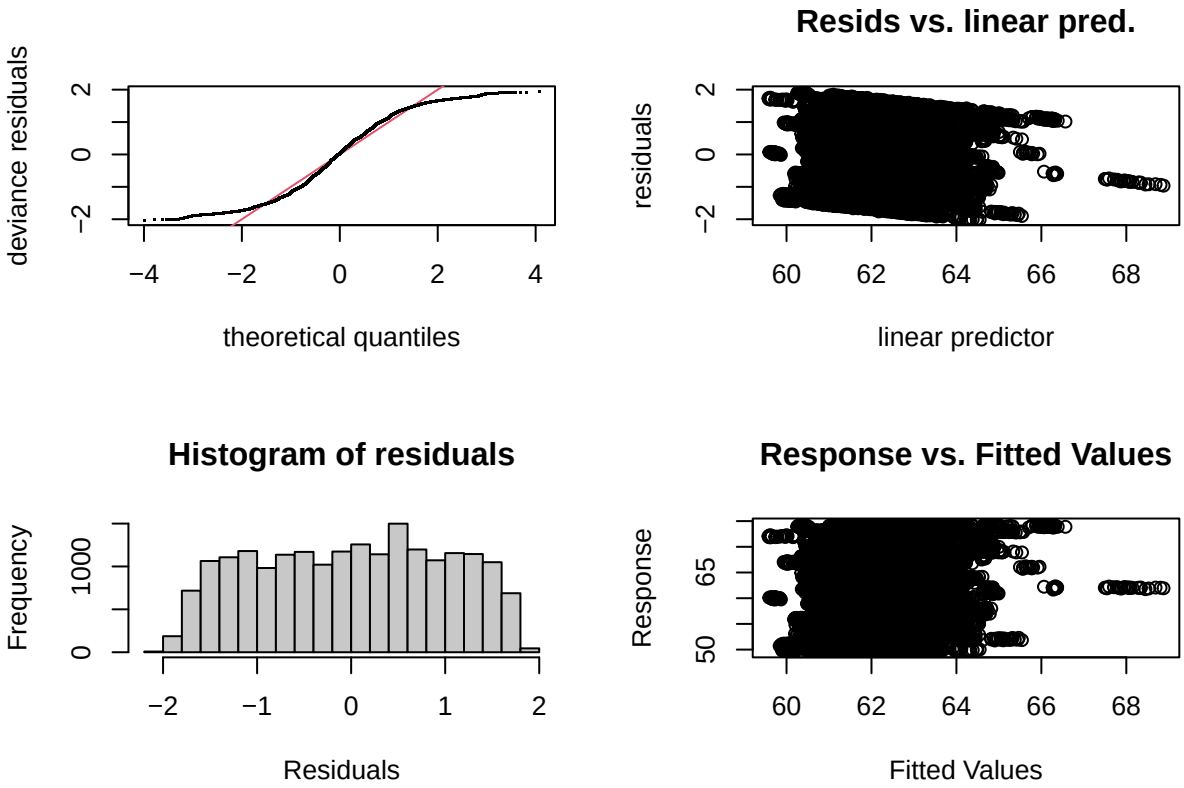
```
##                   df       BIC
## m.final      6.00000  136225.5
## m.final.inv  6.00000 -192625.3
## gam_fit     31.43349  136069.0
```

Student T Distribution GAM fit for better tail performance

```r
gam_t <- gam(
  Resting_BPM ~
    s(Age,  k = 10) +
    s(BMI,  k = 12) +
    s(Water_Intake, k = 10) +
    Gender +
    Workout_Frequency,
  data   = my_data,
  family = scat(link = "identity"),  # Student-t errors
  method = "REML"
)

summary(gam_t)
```

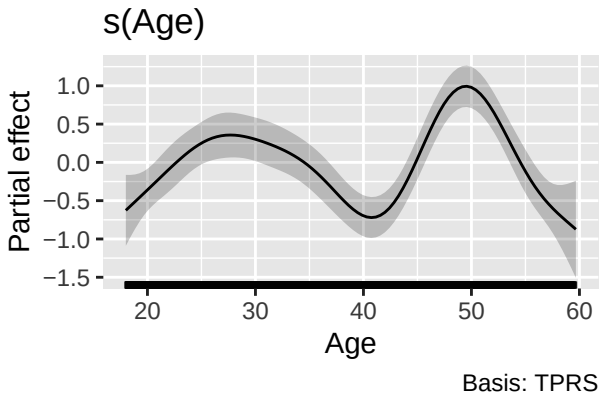
```
##
## Family: Scaled t(Inf,7.212)
## Link function: identity
##
## Formula:
## Resting_BPM ~ s(Age, k = 10) + s(BMI, k = 12) + s(Water_Intake,
##     k = 10) + Gender + Workout_Frequency
##
## Parametric coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       62.71797    0.22000 285.082  < 2e-16 ***
## GenderMale         0.05642    0.10207   0.553  0.58043
## Workout_Frequency -0.16582    0.06264  -2.647  0.00811 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                     edf Ref.df Chi.sq p-value
## s(Age)            7.439  8.400  100.3  <2e-16 ***
## s(BMI)           10.392 10.902  184.0  <2e-16 ***
## s(Water_Intake)   8.525  8.936  149.8  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0211   Deviance explained = 2.25%
## -REML =  67951  Scale est. = 1         n = 20000
```

```r
# Basic diagnostic checks
gam.check(gam_t)
```

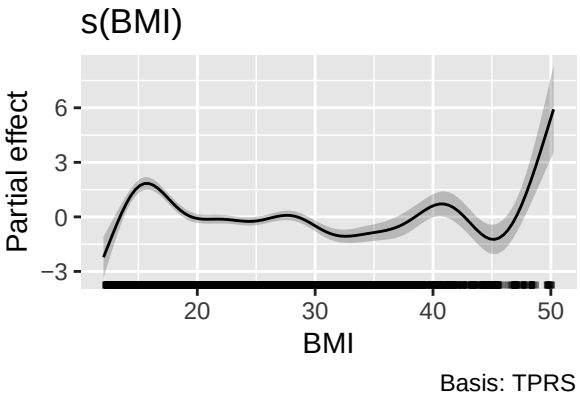



```
##
## Method: REML   Optimizer: outer newton
## full convergence after 8 iterations.
## Gradient range [-0.03130561,0.09574677]
## (score 67950.6 & scale 1).
## Hessian positive definite, eigenvalue range [0.03130509,39929.34].
## Model rank =  32 / 32
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##                   k'   edf k-index p-value
## s(Age)          9.00  7.44    0.96  <2e-16 ***
## s(BMI)         11.00 10.39    0.95  <2e-16 ***
## s(Water_Intake)  9.00  8.53    1.01     0.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

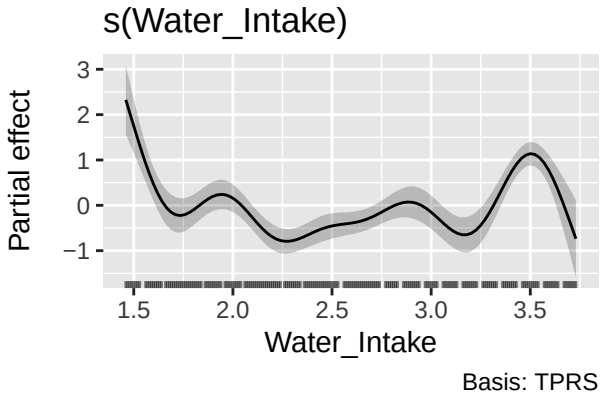
```r
draw(gam_t,
     residuals = FALSE,      # or TRUE if you want them
     scales    = "free")   # each smooth gets its own y-scale
```



Basis: TPRS



Basis: TPRS



Basis: TPRS

AIC Comparison

```
AIC(m.final, m.final.inv , gam_fit, gam_t)
```

```
##                   df       AIC
## m.final      6.00000  136178.0
## m.final.inv  6.00000 -192672.7
## gam_fit     31.43349  135820.6
## gam_t       32.43384  135822.6
```

Acknowledgments

Thank you for reading!

References

[1] Healthdirect Australia, "Resting heart rate – definition, impacting factors & how to check it," , 2024. URL `https://www.healthdirect.gov.au/resting-heart-rate`, authority: Healthdirect Australia Limited.

[2] Quer, G., Gouda, P., Galarnyk, M., Topol, E. J., and Steinhubl, S. R., "Inter- and intraindividual variability in daily resting heart rate and its associations with age, sex, sleep, BMI, and time of year: Retrospective, longitudinal cohort study of 92,457 adults," *PLOS ONE*, Vol. 15, No. 2, 2020, p. e0227709. https://doi.org/10.1371/journal.pone.0227709, URL https://dx.plos.org/10.1371/journal.pone.0227709.

[3] Valentini, M., and Parati, G., "Variables Influencing Heart Rate," *Progress in Cardiovascular Diseases*, Vol. 52, No. 1, 2009, pp. 11–19. https://doi.org/10.1016/j.pcad.2009.05.004, URL https://linkinghub.elsevier.com/retrieve/pii/S0033062009000334.

[4] Ehrenwald, M., Wasserman, A., Shenhar-Tsarfaty, S., Zeltser, D., Friedensohn, L., Shapira, I., Berliner, S., and Rogowski, O., "Exercise capacity and body mass index - important predictors of change in resting heart rate," *BMC Cardiovascular Disorders*, Vol. 19, No. 1, 2019, p. 307. https://doi.org/10.1186/s12872-019-01286-2, URL https://bmccardiovascdisord.biomedcentral.com/articles/10.1186/s12872-019-01286-2.

[5] Alexander, J., Sovakova, M., and Rena, G., "Factors affecting resting heart rate in free-living healthy humans," *DIGITAL HEALTH*, Vol. 8, 2022, p. 205520762211290. https://doi.org/10.1177/20552076221129075, URL http://journals.sagepub.com/doi/10.1177/20552076221129075.

[6] jockeroika), O. E. K. u., "Life style data," , 2025. URL https://www.kaggle.com/datasets/jockeroika/life-style-data.

[7] QCBS R Workshop Series Team, "GAM model checking," , 2023. URL https://r.qcbs.ca/workshop08/book-en/gam-model-checking.html, publication title: QCBS R workshop series: Generalized additive models in R.

[8] Knief, U., and Forstmeier, W., "Violating the normality assumption may be the lesser of two evils," *Behavior Research Methods*, Vol. 53, No. 6, 2021, pp. 2576–2590. https://doi.org/10.3758/s13428-021-01587-5, URL https://link.springer.com/10.3758/s13428-021-01587-5.

[9] Yamaguchi, J., Hozawa, A., Ohkubo, T., Kikuya, M., Ugajin, T., Ohmori, K., Hashimoto, J., Hoshi, H., Satoh, H., and Tsuji, I., "Factors Affecting Home-Measured Resting Heart Rate in the General PopulationThe Ohasama Study," *American Journal of Hypertension*, Vol. 18, No. 9, 2005, pp. 1218–1225. https://doi.org/10.1016/j.amjhyper.2005.04.009, URL https://academic.oup.com/ajh/article-lookup/doi/10.1016/j.amjhyper.2005.04.009.

[10] Avram, R., Kuhar, P., Vittinghoff, E., Aschbacher, K., Tison, G., Pletcher, M., Marcus, G., and Olgin, J., "Abstract 15098: Redefining Normal Resting Heart Rate Values Using Big Data," *Circulation*, Vol. 138, No. Suppl_1, 2018, pp. A15098–A15098. https://doi.org/10.1161/circ.138.suppl_1.15098, URL https://doi.org/10.1161/circ.138.suppl_1.15098, publisher: American Heart Association.