Linear Model Summaries

Model 1

Predictors	Estimate	Std. Error	t value	p-value
Intercept	1031.979	119.843	8.611	2.28e-16 ***
GHI	25.182	0.503	50.062	< 2e-16 ***
Temp	-67.900	4.295	-15.809	< 2e-16 ***
WS	11.469	24.194	0.474	0.636

Note. GHI = Global Horizontal Irradiance, Temp = Ambient Temperature (C), WS = Wind Speed (mph). *** indicates statistical significance. Model is fit to predict total output (kWh).

Residual standard error: 629.1 on 361 degrees of freedom
Multiple R-squared: 0.896, Adjusted R-squared: 0.8951
F-statistic: 1037 on 3 and 361 DF, p-value: < 2.2e-16

Model 2

Predictors	Estimate	Std. Error	t value	p-value
Intercept	411.41561	82.09867	5.011	8.48e-07 ***
GHI	29.89769	0.62316	47.977	< 2e-16 ***
GHI:Temp	-0.36312	0.01883	-19.289	< 2e-16 ***

Note. GHI = Global Horizontal Irradiance, Temp = Ambient Temperature (C), : denotes interaction (product of two variables). *** indicates statistical significance. Model is fit to predict total output (kWh).

Residual standard error: 575.6 on 362 degrees of freedom
Multiple R-squared: 0.9127, Adjusted R-squared: 0.9122
F-statistic: 1893 on 2 and 362 DF, p-value: < 2.2e-16

Model 3

Predictors	Estimate	Std. Error	t value	p-value
Intercept	1227.5186	100.0495	12.27	<2e-16 ***
GHI	19.8135	0.4823	41.09	< 2e-16 ***

Note. GHI = Global Horizontal Irradiance. *** indicates statistical significance. Model is fit to predict total output (kWh).

Residual standard error: 818.5 on 363 degrees of freedom
Multiple R-squared: 0.823, Adjusted R-squared: 0.8225
F-statistic: 1688 on 1 and 363 DF, p-value: < 2.2e-16

Takeaways: Basing our decision off of Multiple R-squared (correlation coefficient) and lowest p-values, we can deduce that Model 2 with the highest R^2 of 0.9127 is the best model to use going forward. The coefficients for this model are 411.41561 for intercept, 29.89769 for GHI and -0.36312 for the interaction effect of GHI and Temp. This can be written into a linear equation of y = 411.41561 + 29.89769(GHI) - 0.36312(GHI:Temp). Where y=kWh and GHI:Temp is the product of the two.

Comparing Measured vs Oiko data Expected Values

GHI five number summary stats:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Measured Data	7.199	94.818	156.302	168.314	249.470	523.074
Oiko Data	9.99	117.72	179.22	185.88	260.16	356.25

Temperature five number summary stats:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Measured Data	-20.364	4.822	13.539	13.417	23.132	33.277
Oiko Data	-20.050	4.982	13.315	13.172	22.657	31.510

Wind Speed five number summary stats:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Measured Data	0.000	1.042	1.896	2.343	3.188	10.115
Oiko Data	0.690	2.453	3.270	3.539	4.438	9.180

T-Test Comparing GHI Means of Oiko vs Measured Data

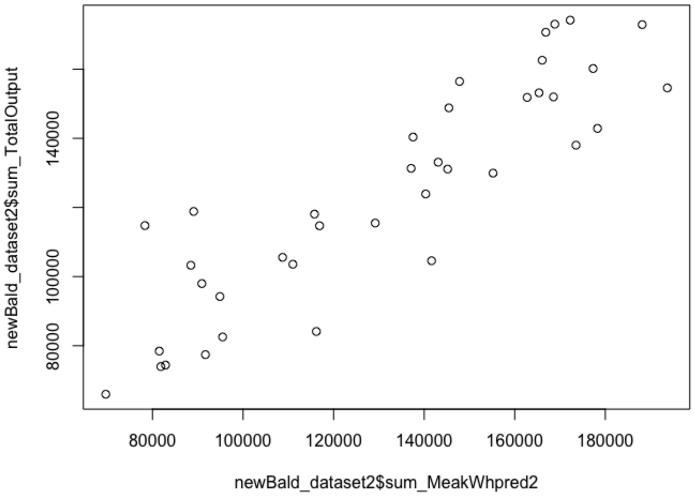
```
Welch Two Sample t-test
data:  dailypredBald$GHI and daily_Bald_Mea$GHI
t = 4.8755, df = 2463.9, p-value = 1.154e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 10.50249 24.63445
sample estimates:
mean of x mean of y
 185.8821  168.3137
```

Takeaways: Of the three variables measured (GHI, Temp and WS) GHI and Wind speed appear to be significantly different between the Oiko and Measured datasets. However, our model of interest does not include wind speed in it, so GHI is what we will focus on. The two sample t-test of GHI from Oiko and measured shows that with a t-value of 4.8755 and corresponding p-value near 0, GHI is significantly higher in the Oiko dataset (mean = 185.8821) than the Measured dataset (mean = 168.3137). This is key information as a higher GHI value will lead to higher predicted kWh values when used in our linear model equation. We can therefore expect Oiko predicted/expected kWh values to be higher than measured data predicted kWh values on average.

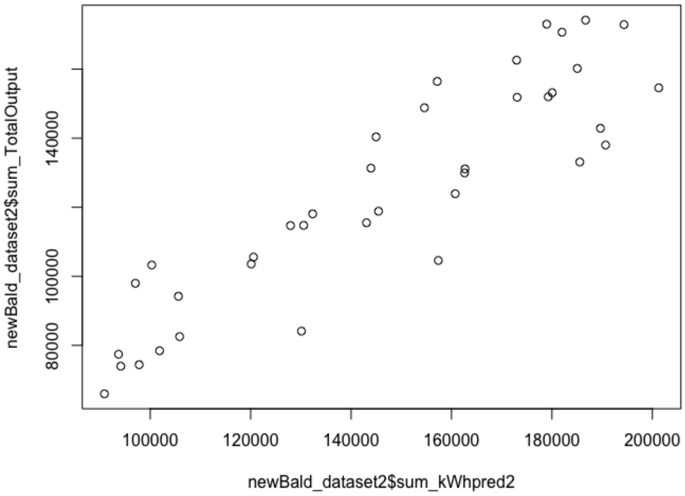
Summary Stats on Predicted vs Observed Values (Using Model 2)

	ER Mean (monthly)	kWh mean (monthly)	RMSE	R
Measured Data	0.9565314	132039.6	17894.59	0.8921388
Oiko Data	0.8475318	146871.8	26913.82	0.89967

Measured data expected vs Observed (kWh)



Oiko expected vs Observed (kWh)



Note. newBald_dataset2$sum_TotalOutput = Observed/Actual Monthly sum (kWh), newBald_dataset2$sum_MeakWh = Monthly sum predicted kWh from measured data, newBald_dataset2$sum_kWhpred2 = Monthly sum predicted kWh from Oiko data.

Takeaways: For starters, like we predicted on the previous page, the monthly average kWh expected value from the Oiko dataset (146871.8) is higher than the monthly average kWh expected value from the measured dataset (132039.6). This translates into the Oiko dataset having a smaller monthly average Energy ratio (0.8475318) than the measured dataset (0.9565314). The RMSE statistic measures the square root of the mean of the differences between expected and observed values. The Oiko dataset produced a RMSE of 26913.82, while the measured dataset had a much lower RMSE of 17894.59. This means there is a lower average difference between expected kWh values from the measured weather dataset and observed kWh values than the Oiko dataset produced. We also have correlation coefficients for both sets of expected values against observed values. This statistic measures the amount of variation that can be explained by the variable, in this case expected kWh. The Oiko dataset expected values had a correlation coefficient of 0.89967, while the measured dataset expected values had a correlation coefficient of 0.89214. Both of these values are very high as we'd expect and very close to one another. While both statistics are important, for our purpose I believe RMSE to be more vital and more aligned with Energy Ratios, so from this I would say the measured weather data makes a better set of predicted kWh values than Oiko.