Predicting Solar Power Generation

Alexander McIntosh

7/15/2022

Contents

1 Introduction

As the unintended consequences of traditional energy collection become untenable, the world is turning to alternative, renewable energy sources. Solar power has become a particularly popular method of energy extraction. However, a notable drawback of solar power is its lack of day-to-day consistency. Inconsistency is a significant challenge for interfacing with an external power grid.

In the analysis to follow, using only the data collected at the solar power plant, the power output of the plant will be predicted based only on data available to the plant. First, suspicious data are examined for usability. Then, the clean data are explored for patterns, correlation, and causality. Finally, four models are engineered using linear regression, ARIMA, random forest machine learning, and dimension reduction through principle component analysis. The success of each model is measured by the model's ability to predict three days worth of data, separated before the models' construction.

For brevity and readability, many of the transformations used on the data to produce graphs are not included in this report. For more information about how the graphs and results were generated, see http://www.github.com/mcintalmo/solar-plant-power-generation . The link includes the R Markdown that generated this report and the notebook used for exploration.

To explore the raw data, see *Solar Power Generation Data* from user Ani Kanal on Kaggle. https://www.kaggle.com/datasets/ef9660b4985471a8797501c8970009f36c5b3515213e2676cf40f540f0100e54 Or use command

```
kaggle datasets download -d anikannal/solar-power-generation-data
```

Table 1: Power Generation Data Dimensions

Observations	Variables
136476	7

Table 2: First 5 Observations Of Power Generation Data

date_time	plant_id	source_key	dc_power	ac_power	daily_yield	total_yield
2020-05-15	4135001	1BY6WEcLGh8j5v7	0	0	0	6259559
2020-05-15	4135001	1IF53ai7Xc0U56Y	0	0	0	6183645
2020-05-15	4135001	3PZuoBAID5Wc2HD	0	0	0	6987759
2020-05-15	4135001	7JYdWkrLSPkdwr4	0	0	0	7602960
2020-05-15	4135001	McdE0feGgRqW7Ca	0	0	0	7158964

2 Methods and Analysis

This analysis was conducted entirely within R. Packages included tidyverse and broom for data organization and transformation. Lubridate and hms allowed for time stamp transformation. Tsibble, forecast, fable, and feasts provided the backbone of analysis of the data as a time series and use of the ARIMA model. Ranger, randomForest, pls, and caret all contributed functions for developing a machine learning model. kableExtra and bookdown were used only in the report to generate more visually appealing tables and visualizations.

2.1 Data Collection

The data for this analysis was collected by two solar plants in India. The first plant, id: 4135001, is near Gandikotta, Andhra and the second plant, id: 4136001, is near Nasik, Maharashtra. The data were collected at 15 minute intervals for 34 days. 22 inverter sensors and one weather sensor at each plant brought the total to 46 sensors collecting data.

The inverter sensors collected the DC and AC output of the group of solar panels it was monitoring. Additionally, the inverters tallied the daily yield in DC output from midnight to midnight and maintained the total yield the sensor had observed. Total yield increased until a sensor was replaced, at which point it would begin again at 0.

The weather sensors recorded the ambient temperature, the temperature of the module, and the irradiation level. Owing to the sensors' reading of 0 at night, it is assumed any irradiation is owed entirely to the sun's radiation.

2.2 Data Cleaning

2.2.1 The Raw Data

The data are spread out over four comma separated files; two power generation files and two weather sensor files, one each for each plant. First, the data are read into memory, using lubridate to parse the time stamp, coercing the plant_id to a character, and renaming variables for usability. Note, one of the power generation files has a different date format than the other three files. This suggests human input or conversion was involved in the collection of the data.

2.2.2 Loading the Data

The generation data consist of 136476 observations of 7 variables: The time stamp, the plant id number, the inverter key, the measured DC power from the panels, the measured AC after inversion, the daily yield of power, and the total yield that the sensor has ever recorded.

Table 3: Power Generation Source Keys Separated by Plant ID

4135001	4136001
1BY6WEcLGh8j5v7	4UPUqMRk7TRMgml
1IF53ai7Xc0U56Y	81aHJ1q11NBPMrL
3PZuoBAID5Wc2HD	9kRcWv60rDACzjR
7JYdWkrLSPkdwr4	Et9kgGMDl729KT4
adLQvlD726eNBSB	IQ2d7wF4YD8zU1Q
bvBOhCH3iADSZry	LlT2YUhhzqhg5Sw
iCRJl6heRkivqQ3	LYwnQax7tkwH5Cb
ih0vzX44oOqAx2f	mqwcsP2rE7J0TFp
McdE0feGgRqW7Ca	Mx2yZCDsyf6DPfv
pkci93gMrogZuBj	NgDl19wMapZy17u
rGa61gmuvPhdLxV	oZ35aAeoifZaQzV
sjndEbLyjtCKgGv	oZZkBaNadn6DNKz
uHbuxQJl8lW7ozc	PeE6FRyGXUgsRhN
VHMLBKoKgIrUVDU	q49J1IKaHRwDQnt
wCURE6d3bPkepu2	Qf4GUc1pJu5T6c6
WRmjgnKYAwPKWDb	Quc1TzYxW2pYoWX
YxYtjZvoooNbGkE	rrq4fwE8jgrTyWY
z9Y9gH1T5YWrNuG	V94E5Ben1TlhnDV
zBIq5rxdHJRwDNY	vOuJvMaM2sgwLmb
ZnxXDlPa8U1GXgE	WcxssY2VbP4hApt
ZoEaEvLYb1n2sOq	xMbIugepa2P7lBB
zVJPv84UY57bAof	xoJJ8DcxJEcupym

Table 4: Weather Data Dimensions

Observations	Variables
6441	6

Table 5: First 5 Observations Of Weather Data

date_time	plant_id	source_key	ambient_temperature	module_temperature	irradiation
2020-05-15 00:00:00	4135001	HmiyD2TTLFNqkNe	25.2	22.9	0
2020-05-15 00:15:00	4135001	HmiyD2TTLFNqkNe	25.1	22.8	0
2020-05-15 00:30:00	4135001	HmiyD2TTLFNqkNe	24.9	22.6	0
2020-05-15 00:45:00	4135001	HmiyD2TTLFNqkNe	24.8	22.4	0
2020-05-15 01:00:00	4135001	HmiyD2TTLFNqkNe	24.6	22.2	0

Table 6: Weather Data Source Keys Separated by Plant ID

4135001	4136001
HmiyD2TTLFNqkNe	iq8k7ZNt4Mwm3w0

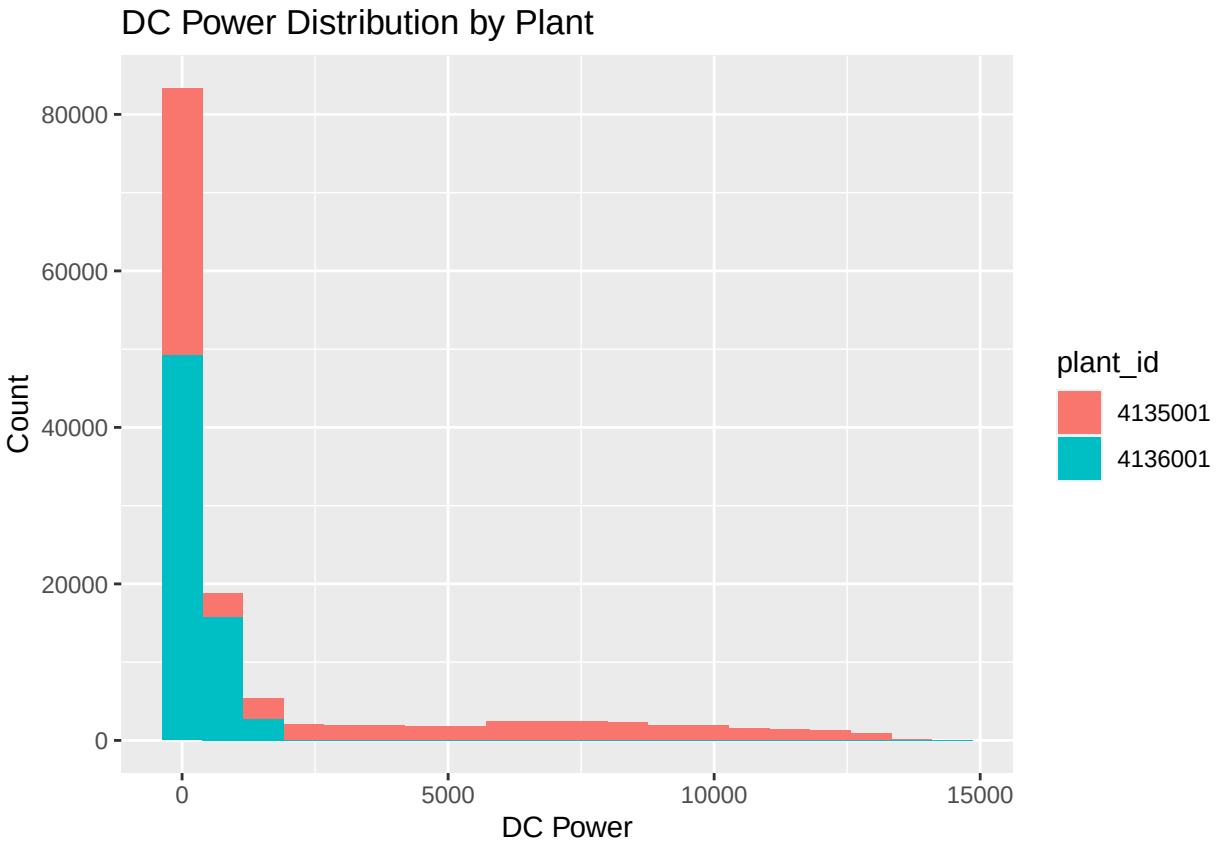
Table 7: DC Power Summary by Plant

plant_id	min_dc	max_dc	median_dc	mean_dc
4135001	0	14471	429	3147
4136001	0	1421	0	247

The weather data consist of only 6441 observations of 6 variables: The time stamp of the observations at 15 minute intervals, the plant id at which the observation was taken, the weather sensor key (identical accross the plant), the ambient temperature in Celsius, the module temperature in Celsius, and the irradiation. The difference in the number of observations owes to the number of sensors. Each plant has 22 units recording power generation data, and 1 sensor recording weather data.

2.2.3 Sanity Checks

First, let us take a look at a summary of the generation data by plant in figure @ref(fig:dc_power_distribution_by_plant).



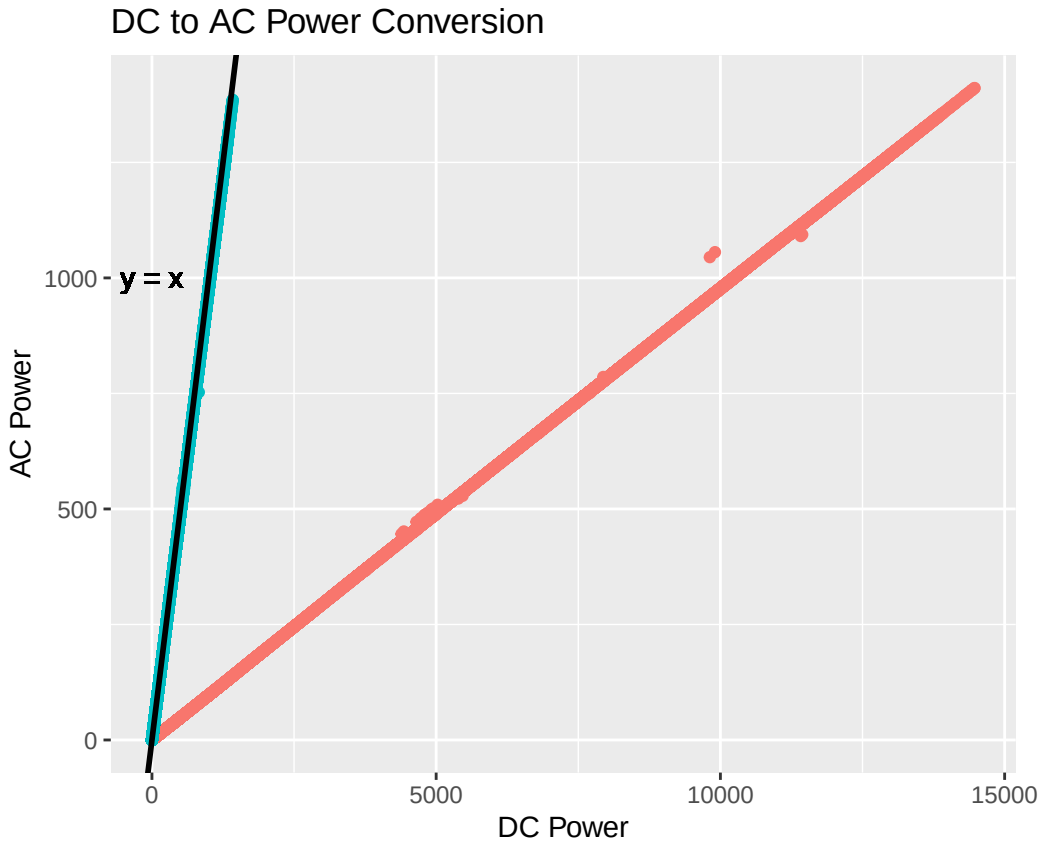
The max and mean values are different by about a factor of 10. By the histogram, it may be seen that the distribution of DC Power observations is considerably more spread out for plant 4135001 than 4136001. Over half of the recorded observations at plant 4136001 are 0. That is not unreasonable, if we expect the sun to be down for more than half of the day.

However, these data were collected in May in India, which resides in the Northern hemisphere. It would be expected that the sun is out for more than half of the observations.

To better understand what might be happening, let us compare the DC output of the solar panels to the AC out-

plant_id	date_time	generation_source	dc_power	ac_power	daily_yield	total_yield	weather_source
4135001	2020-05-15	1BY6WEcLGh8j5v7	0	0	0	6259559	HmiyD2TTLFNqkN
4135001	2020-05-15	1IF53ai7Xc0U56Y	0	0	0	6183645	HmiyD2TTLFNqkN
4135001	2020-05-15	3PZuoBAID5Wc2HD	0	0	0	6987759	HmiyD2TTLFNqkN
4135001	2020-05-15	7JYdWkrLSPkdwr4	0	0	0	7602960	HmiyD2TTLFNqkN
4135001	2020-05-15	adLQvlD726eNBSB	0	0	0	6271355	HmiyD2TTLFNqkN

Observations	Variables
143616	11



put after conversion.
Plant 4136001 is reporting that the conversion rate from DC power to AC power is about %100. However, 4135001 is reporting a 90% loss of power during conversion, despite collecting 10 times as much power.

Even the worst DC to AC adapters are about 80% effective.

This suggests that a conversion error changed the DC Power variable from plant 4135001 by a factor of 10.

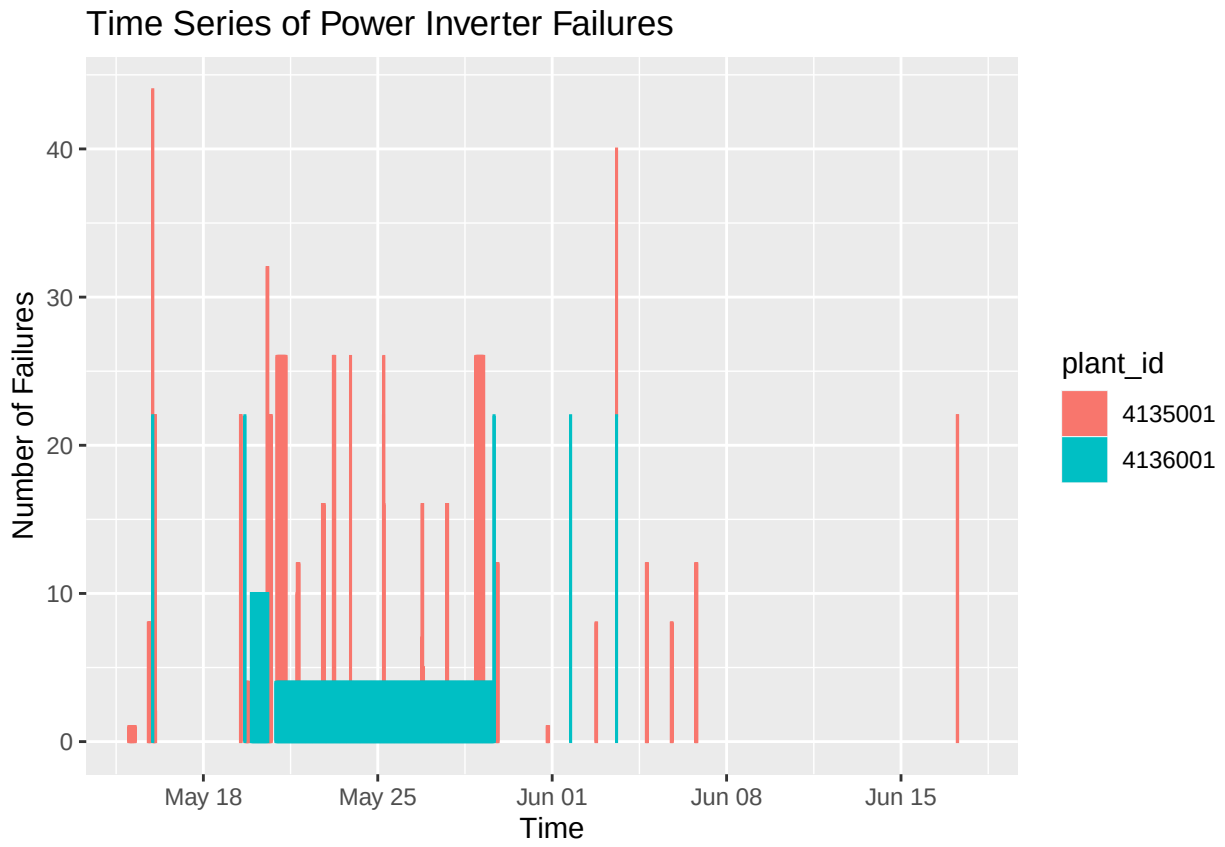
The conversion error is assumed for the remainder of the analysis, and the DC power values for plant 4135001 are divided by 10.

2.2.4 Missing Observations

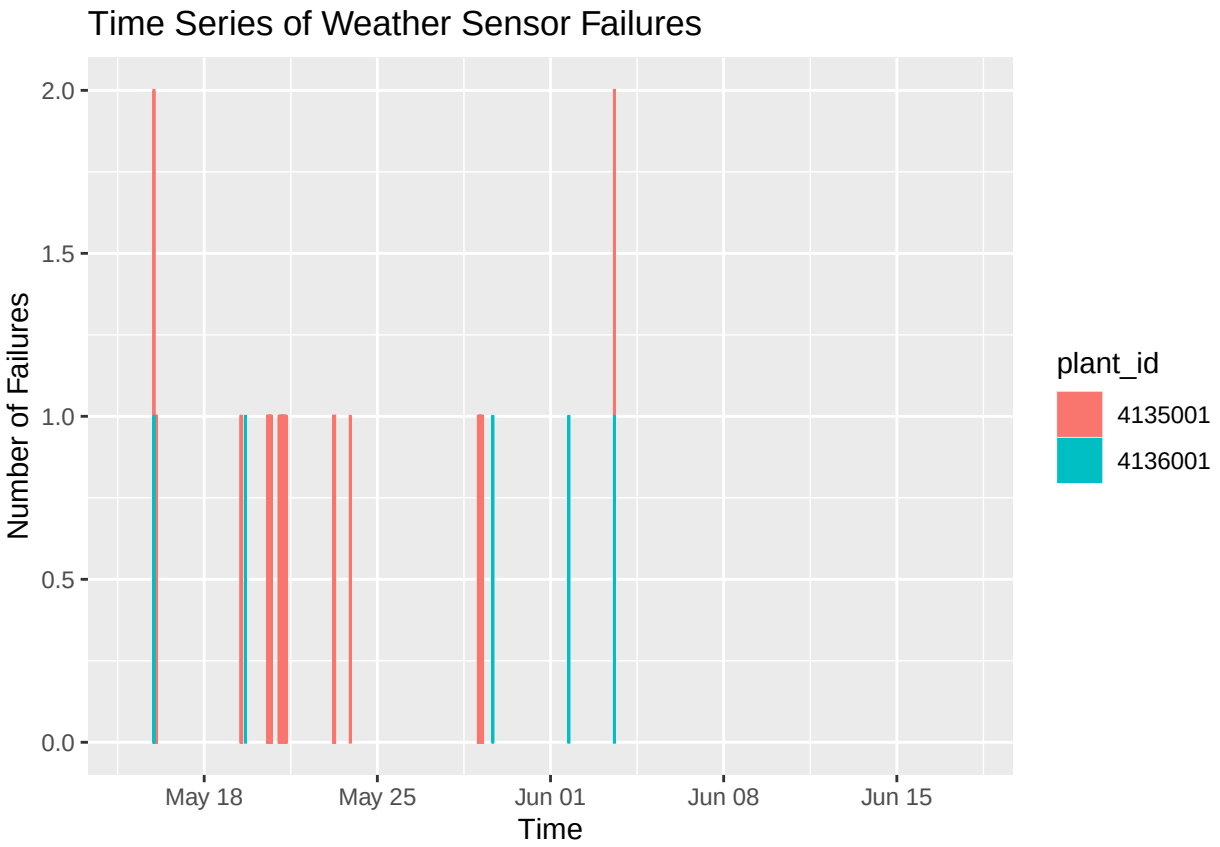
As noted previously, there were 136476 power generation observations made. However, if an observation is made every 15 minutes at 44 sensors for 34 days, we expect to see 143616 observations.

The remaining 7140 observations are missing, and will be considered NA for time series analysis. Can a pattern to the missing values be found? To start, we will fill in the missing time values and combine the weather data.

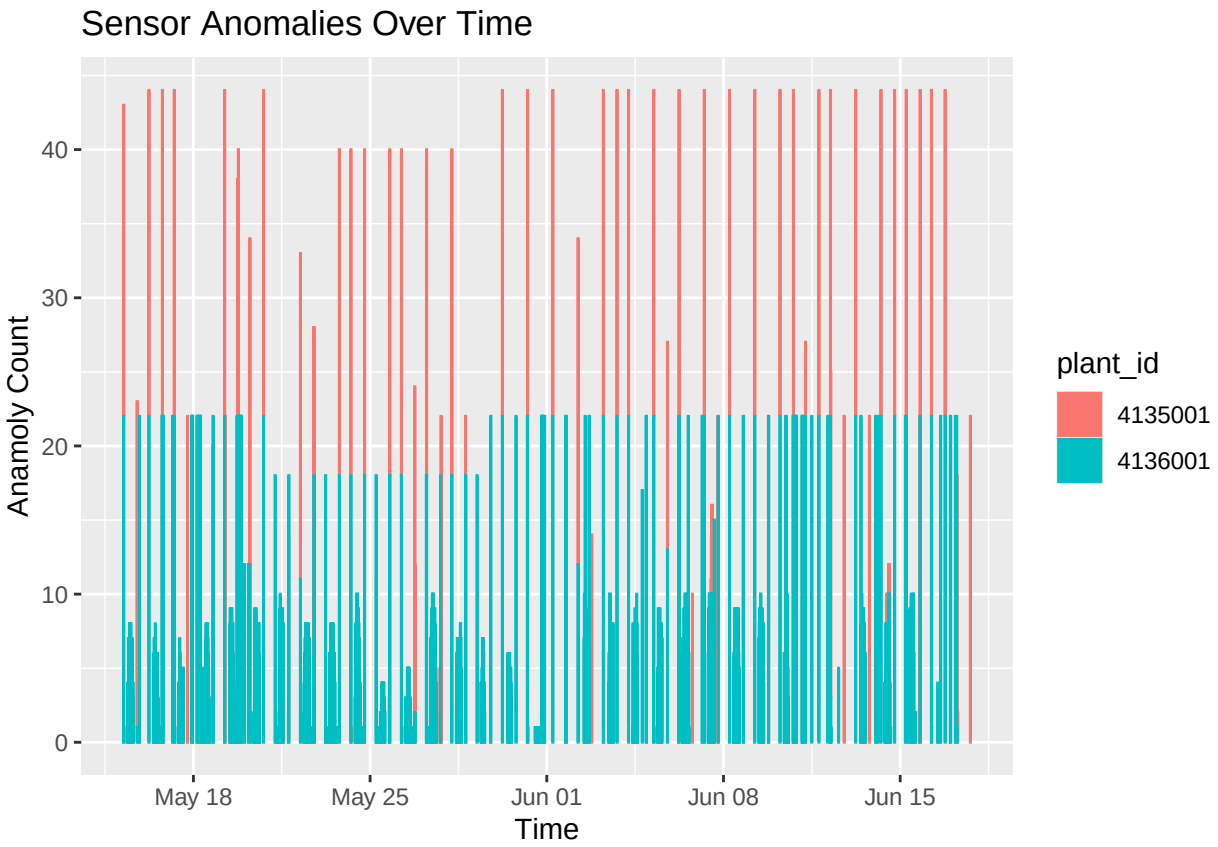
143616 observations of 11 distinct variables, all times are now accounted for.



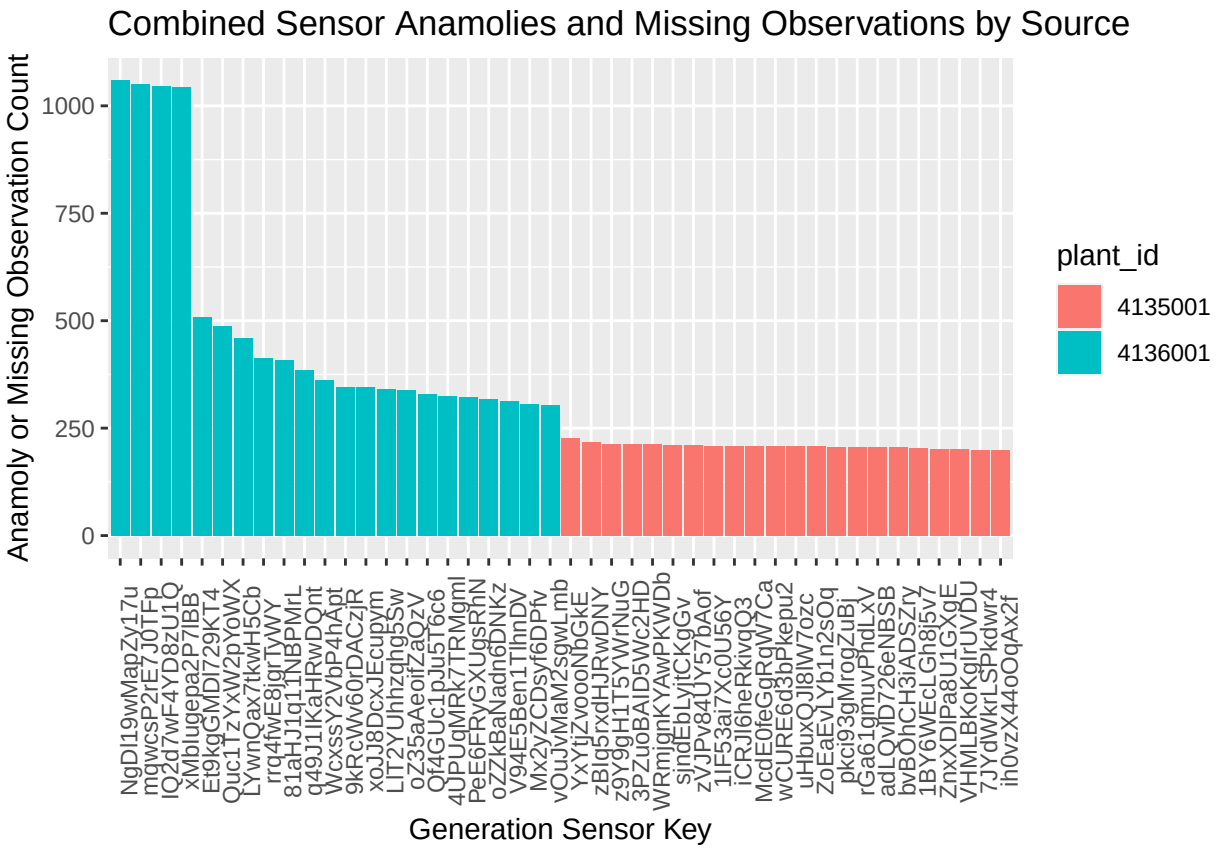
The above graph shows the number of missing values from power sensors over time. Notably, missing values at plant 4135001 seem to go across all 22 sensors. Additionally, For much of the last week of may, plant 4136001 had about 4 power generation sensors that were offline.



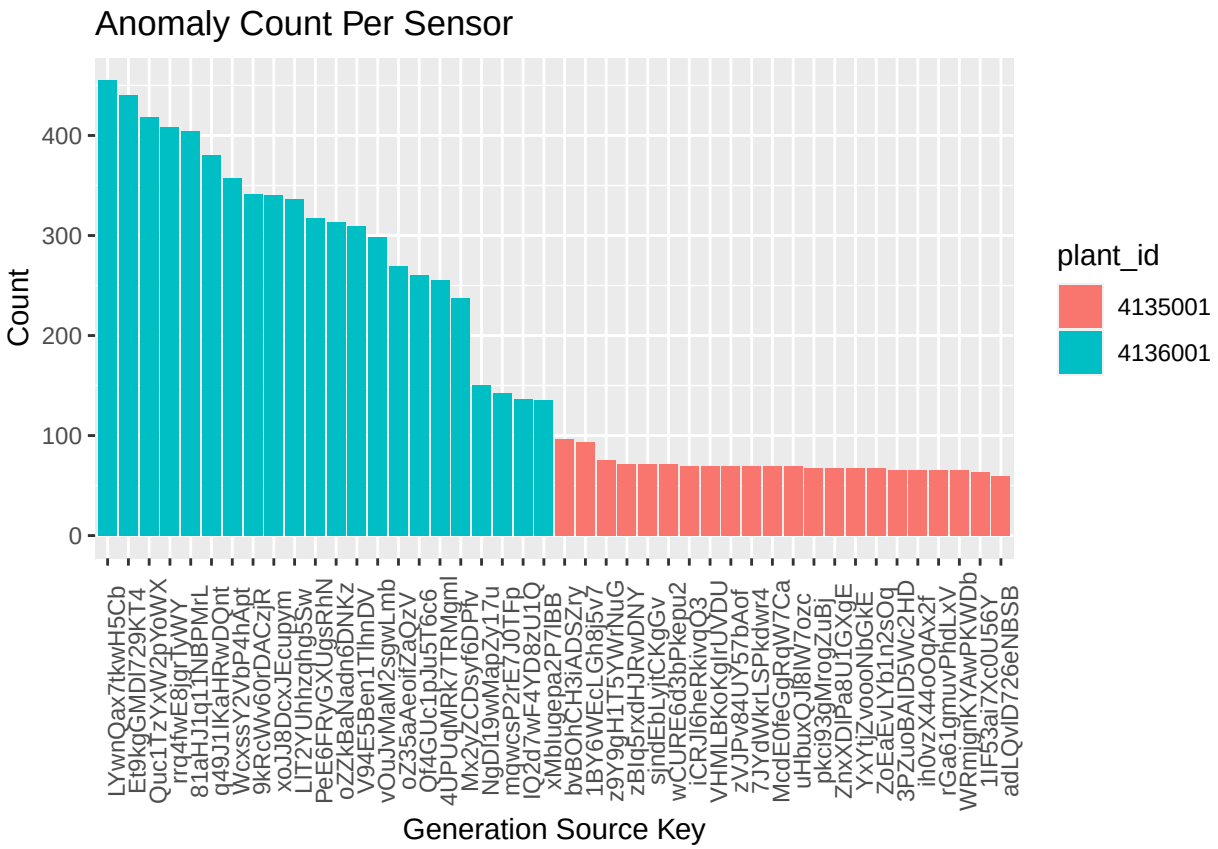
The above graph shows the number of missing values from power sensors over time. Note that many of the missing values of one or both weather sensors align with missing values of the generation sensors as well. This might suggest maintenance or a failure of a kind that affected the entire plant.



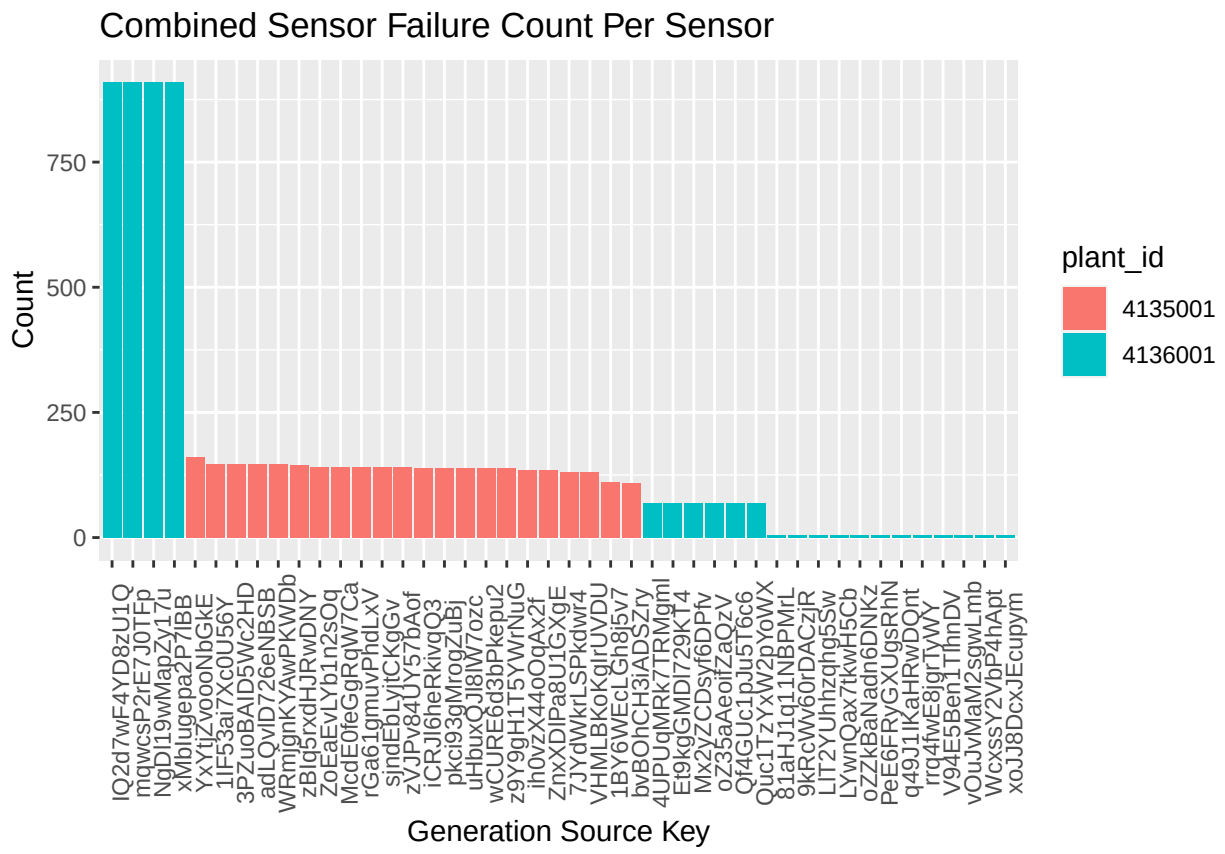
The above graph displays anomalies in the sensor data. Here, an anomaly is defined as an observation in which the irradiation measured by the weather sensor is greater than 0, but no DC power is generated. Without more information, it is hard to tell exactly what is happening in these cases. For example, is it true that the panels are not actually producing DC power, or is it the case simply that the power sensor is not detecting the generated power.

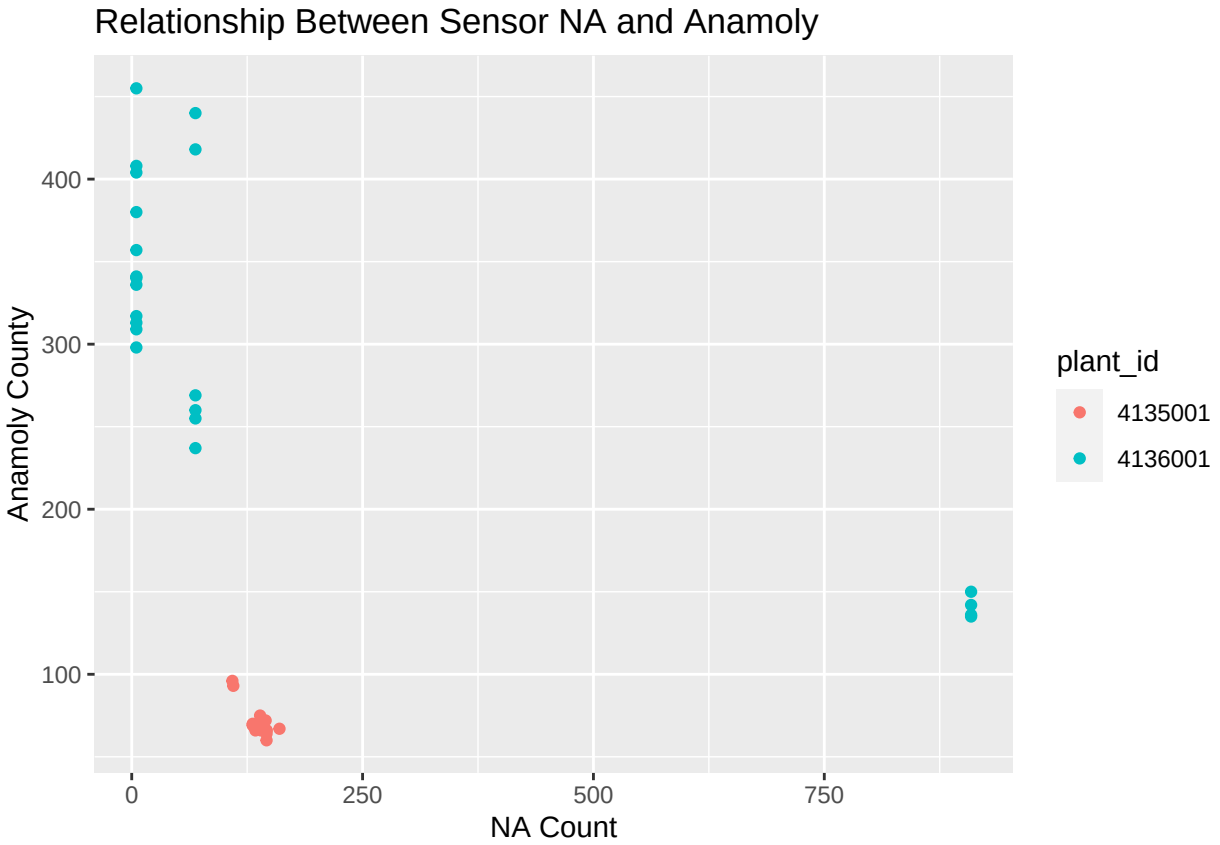


It is time to get more granular. Are there specific sensors that might be contributing disproportionately to missing and anomalous data? The above graph shows the total number of combined anomalies and missing observations broken down for each source. It is clear that plant 4136001 experiences many more problems, with four sources contributing many problems. Even the best 4136001 sensor performs worse than the worst 4135001 sen-


Anomaly Count Per Sensor
Count
400
300
200
100
0
plant_id
4135001
4136001
LYwnQax7tkwH5Cb
Et9kgGMDl729kT4
Quc1TzYxW2pYoWX
rrq4fwE8igrTvWY
81aHJ1q1INBPMrL
q49J1lKaHRwDOnt
WcxssY2VbP4hApt
9kRcWv60rDACzjR
xoJJ8DcxJEcupym
LIT2YUhhzqhg5Sw
PeE6FRyGXUgsRhN
oZZkBaNadn6DNKz
V94E5Ben1TlhnDV
vOuJvMaM2sgwLmb
oZ35aAeoifZaQzV
Of4GUc1pJu5T6c6
4UPUqMRK7TRMgml
Mx2yZCDsyf6DPfv
NgDl19wMapZy17u
mqwcsP2rE7J0TFp
IQ2d7wF4YD8zU1Q
xMblugepa2P7lBB
bvBOhCH3iADSZry
1BY6WEcLGh8j5v7
z9Y9gH1T5YWrNuG
zBIq5rxdHJRwDNY
sJndEbLyjtCKgGv
wCURE6d3bPkepu2
iCRJl6heRkivqQ3
VHMLBKoKgirUVDU
zVJPv84UY57bAof
7JYdWkrLSPkdwr4
McdE0feGgRqW7Ca
uHbuxQJl8lW7ozc
pkci93gMrogZuBj
ZnxXDlPa8U1GXgE
YxYtjZvoooNbGkE
ZoEaEvLYb1n2sOq
3PZuoBAID5Wc2HD
ih0vzX44oOqAx2f
rGa61gmuvPhdLxV
WRmjgnKYAwPKWDb
1IF53ai7Xc0U56Y
adLQvID726eNBSB
Generation Source Key

sor.


Combined Sensor Failure Count Per Sensor
Count
750
500
250
0
IQ2d7wF4YD8zU1Q
mqwcsP2rE7J0TFp
NgDl19wMapZy17u
xMblugepa2P7lBB
YxY1jZvoooNbGkE
1lF53ai7Xc0U56Y
3PZuoBAID5Wc2HD
adLQvlD726eNBSB
WRmjgnKYAwPKWDb
zBlg5rxdHJRwDNY
ZoEaEvLYb1n2sOq
McdE0feGgRqW7Ca
rGa61gmuvPhdLxV
sjndEbLyjtCKgGv
zVJPv84UY57bAof
iCRJl6heRkivgQ3
pkci93gMrogZuBj
uHbuxQJl8lW7ozc
wCURE6d3bPkepu2
z9Y9gH1T5YWrNuG
ih0vzX44oOqAx2f
ZnxXDlPa8U1GXgE
7JYdWkrLSPkdwr4
VHMLBKoKglrUVDU
1BY6WEcLGh8j5v7
bvBOhCH3iADSZry
4UPUqMRk7TRMgml
Et9kgGMDI729kT4
Mx2yZCDsyf6DPfv
oZ35aAeoifZaOzV
Qf4GUc1pJu5T6c6
Quc1TzYxW2pY0WX
81aHJ1q11NBPMrL
9kRcWv60rDACzjR
LIT2YUhhzqhg5Sw
LYwnQax7tkwH5Cb
oZZkBaNadn6DNKz
PeE6FRyGXUgsRhN
q49J1lKaHRwDQnt
rrq4fwE8jgrTyWY
V94E5Ben1TfhnDV
vOuJvMaM2sgwLmb
WcxssY2VbP4hApt
xoJJ8DcxJEcupym
Generation Source Key
plant_id
4135001
4136001



The prior three graphs separate the combined anamoly counts from the missing observation count. The third graph shows the sensor grouping as a result of relating missing counts to anamoly counts. The plant 4136001 cluster shows the healthiest collection of data, while plant 4135001 clusters suggest that less missing data from a set does not necessarily mean that the data is more trustworthy.

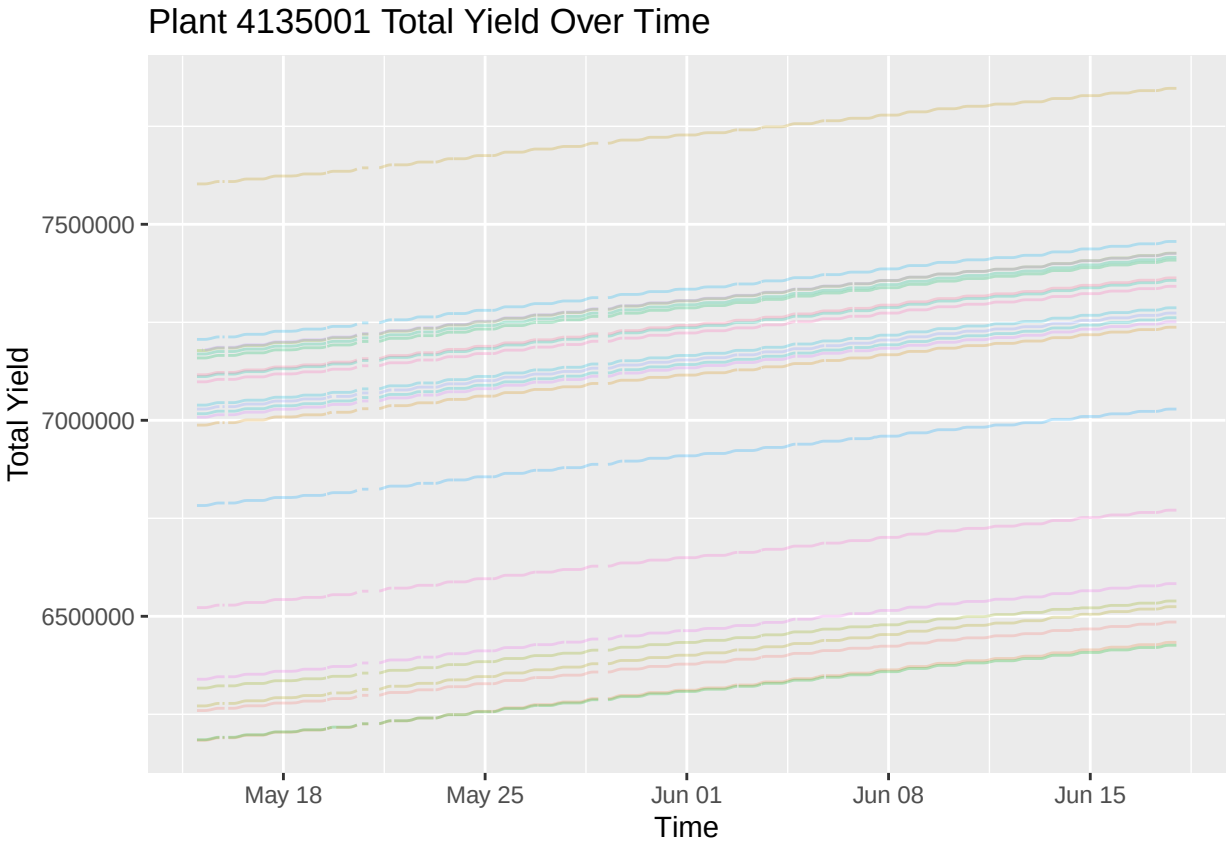
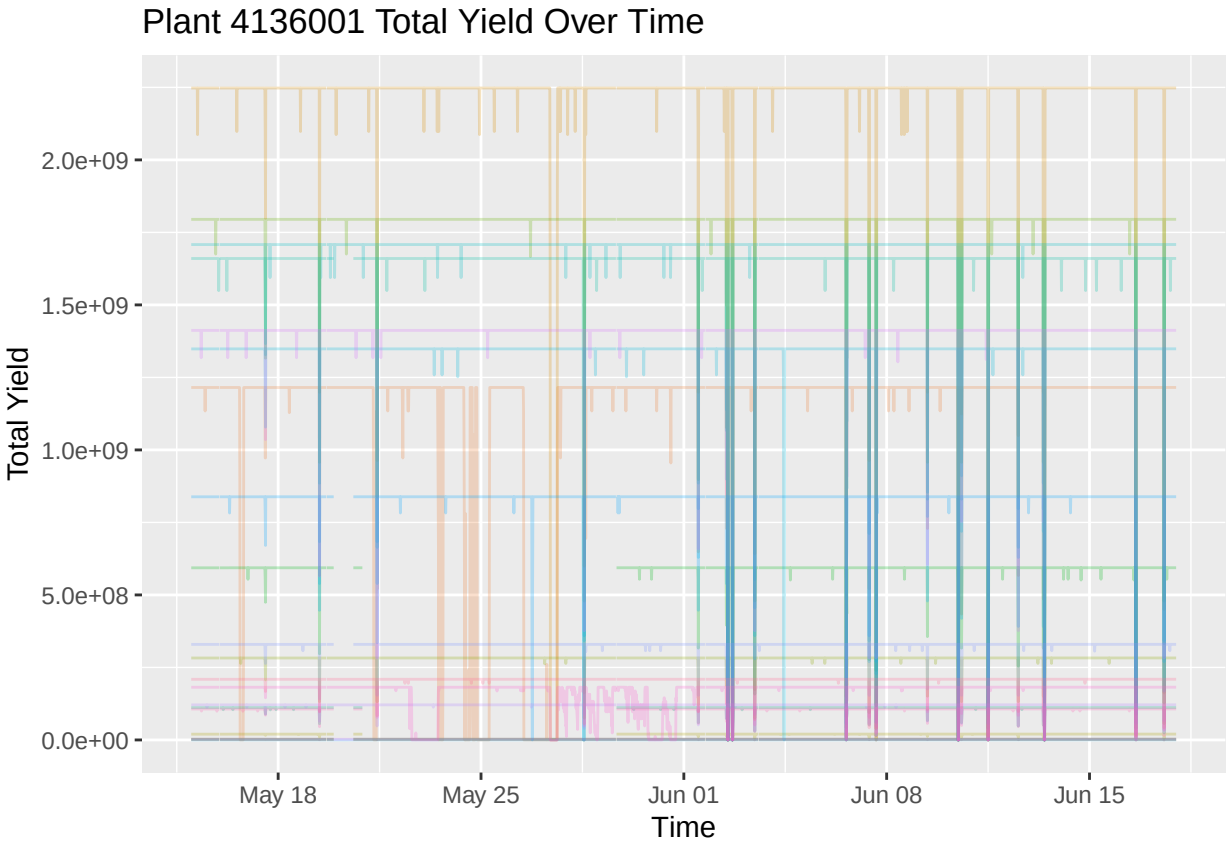

Plant 4135001 Total Yield Over Time
Total Yield
7500000
7000000
6500000
May 18
May 25
Jun 01
Jun 08
Jun 15
Time

Table 8: Summary of Power Generation Data

Variable	min	max	median	mean	sd
dc_power	0	1441	44.9	320	409
ac_power	0	1405	43.4	313	399
daily_yield	0	9163	2559.9	3307	3186



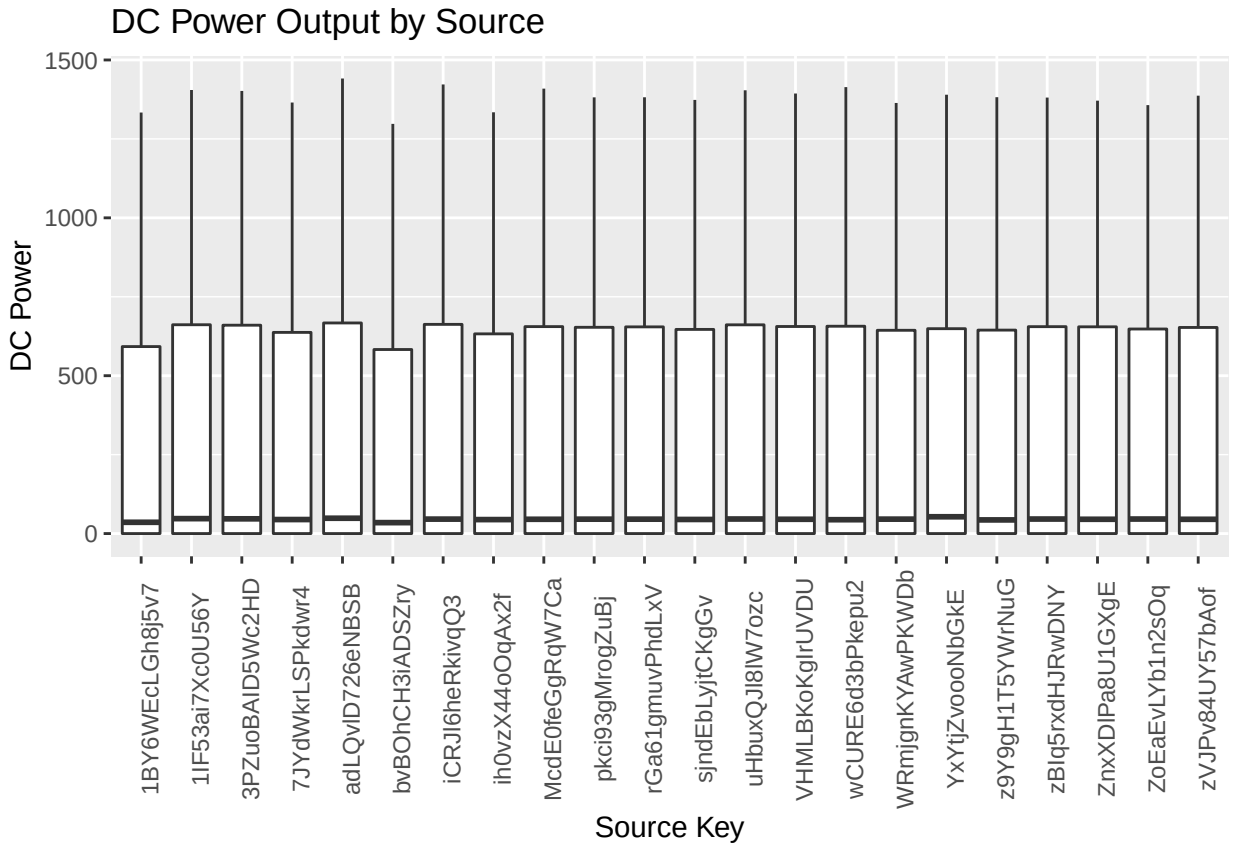
For more evidence that plant 4136001 is better left out of model construction, see the second graph above. Though it is not clear what might cause these readings from the sensors, it suggests frequent problems with the sensors. This, combined with the evidence provided by analyzing the DC power output of the panels, justifies removal of plant 4136001 from the data set for model development.

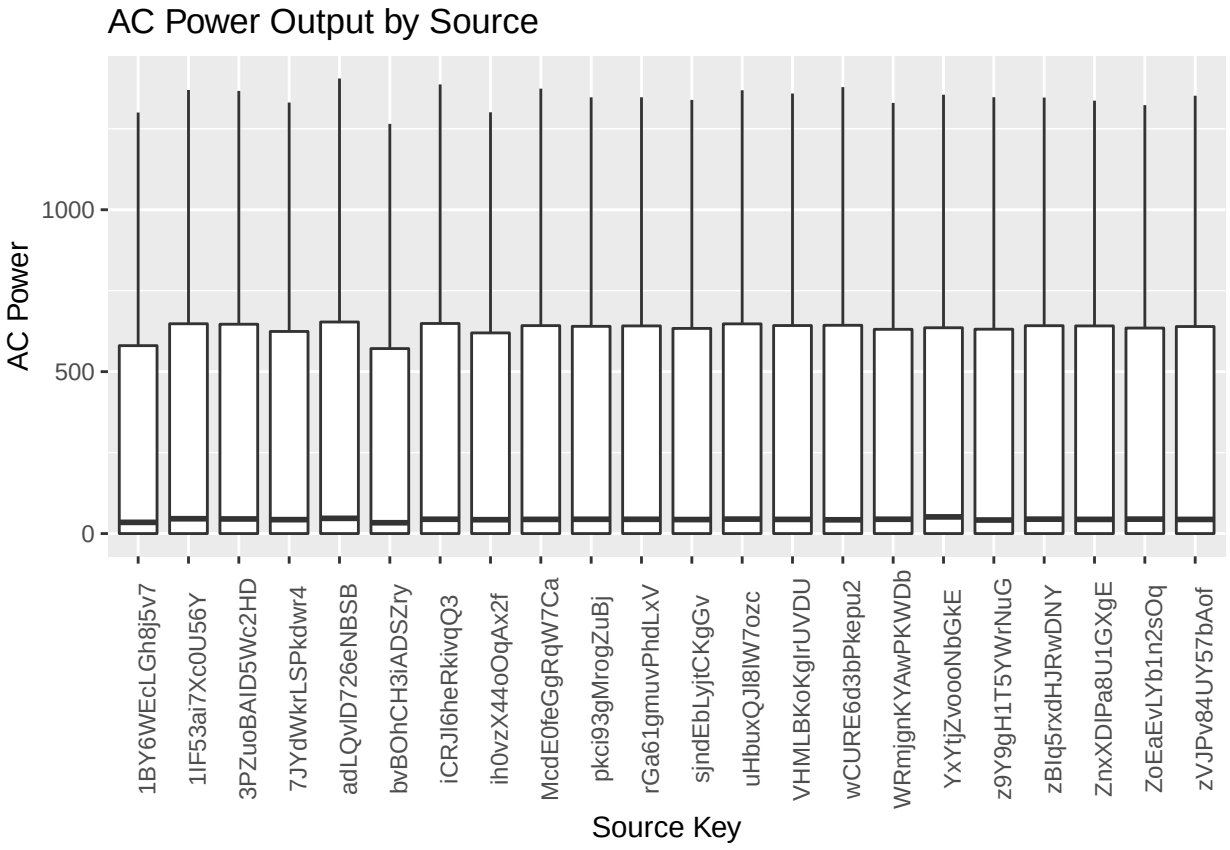
2.3 Separating Validation Data

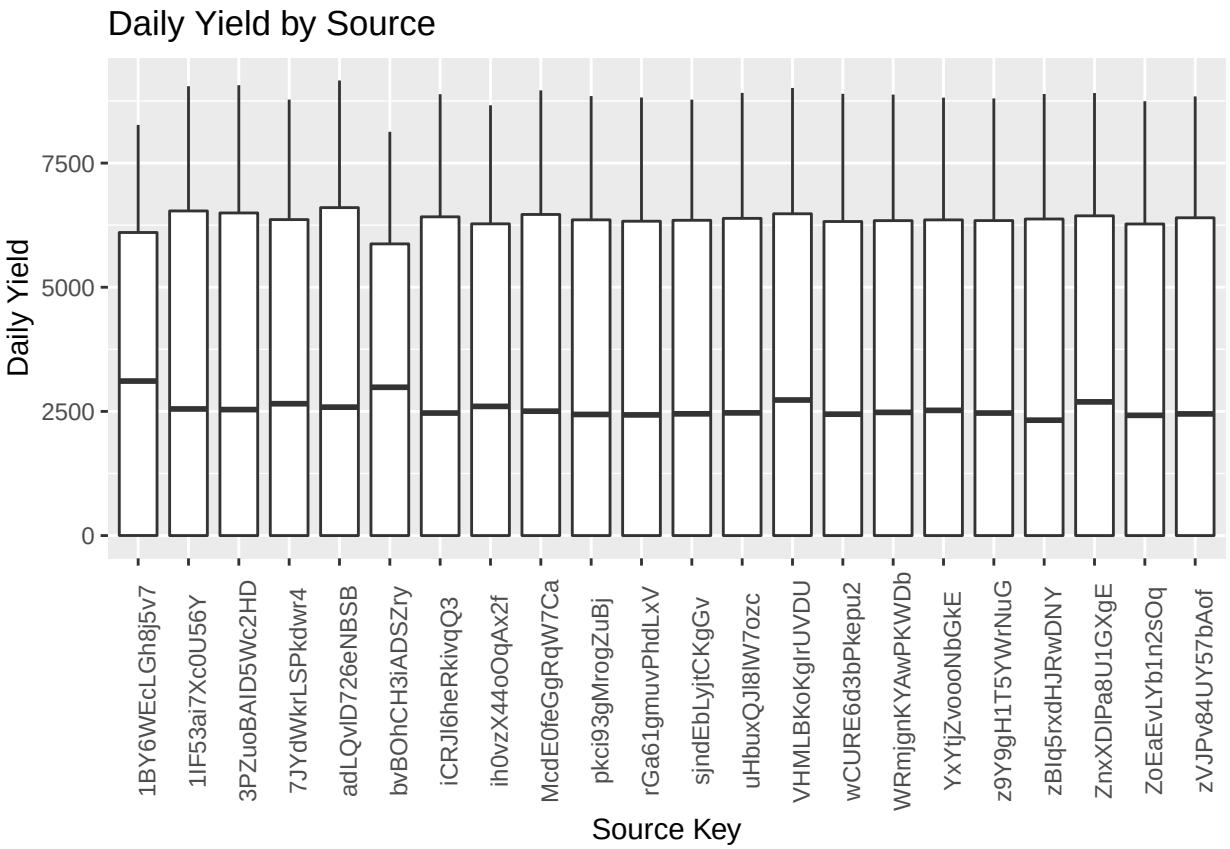
To best determine the overall efficacy of predictive models, the data sets were broken into a validation set and a training set. There were 34 days in the entire data set. So, the last three days were partitioned to be a validation set, against which predictions would only be made after the models were finalized. Later, to build the models themselves, the training data was separated further into a train set and test set, using days 30 - 31 for comparison.

2.4 Data Exploration

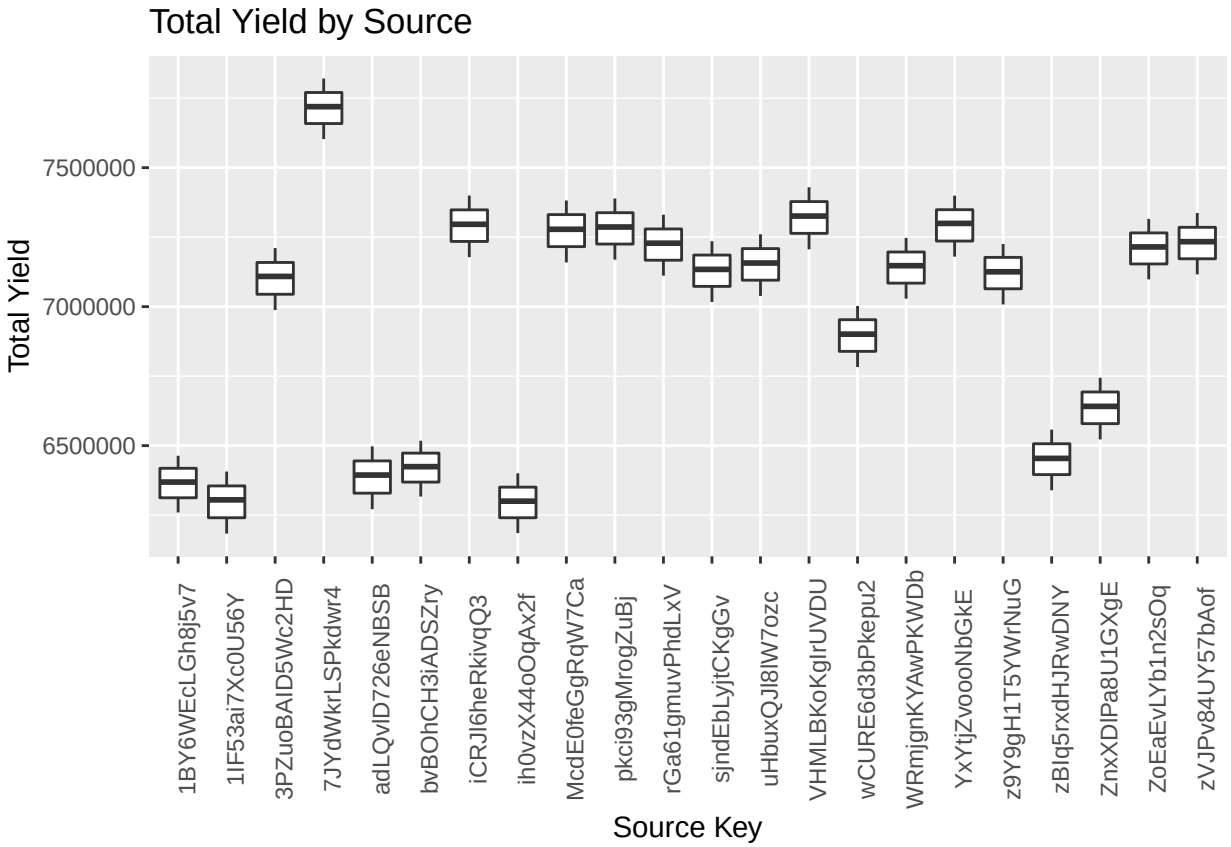
The data has been cleaned. A validation set has been partitioned. Before models could be constructed, the data must be further explored to identify patterns. ### Generation {#generation}

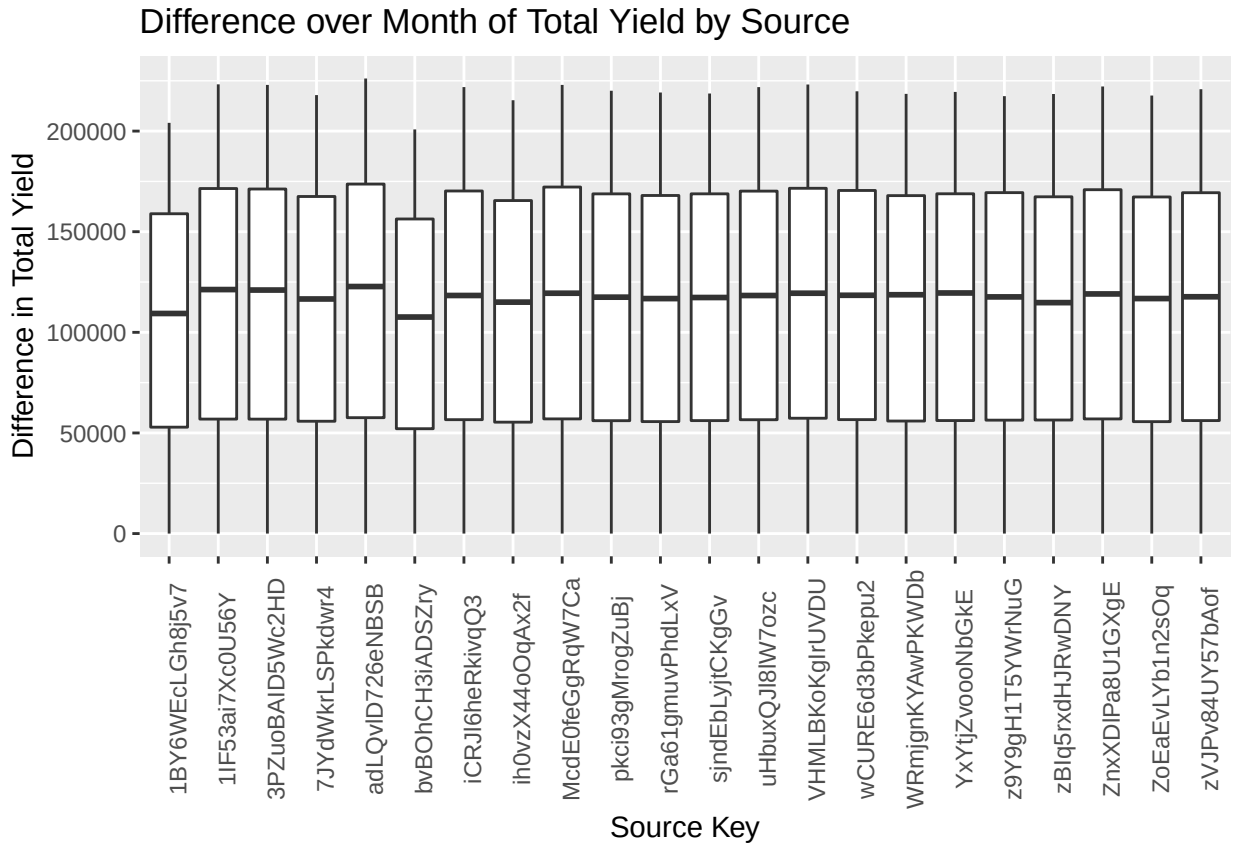

DC Power Output by Source
DC Power
1500
1000
500
0
1BY6WEcLGh8j5v7
1IF53ai7Xc0U56Y
3PZuoBAID5Wc2HD
7JYdWkrLSPkdwr4
adLQvlD726eNBSB
bvBOhCH3iADSZry
iCRJl6heRkivqQ3
ih0vzX44oOqAx2f
McdE0feGgRqW7Ca
pkci93gMrogZuBj
rGa61gmuvPhdLxV
sjndEbLyjtCKgGv
uHbuxQJl8lW7ozc
VHMLBKoKglrUVDU
wCURE6d3bPkepu2
WRmjgnKYAwPKWDb
YxYtjZvoooNbGkE
z9Y9gH1T5YWrNuG
zBIq5rxdHJRwDNY
ZnxXDlPa8U1GXgE
ZoEaEvLYb1n2sOq
zVJPv84UY57bAof
Source Key


AC Power Output by Source
AC Power
1000
500
0
1BY6WEcLGh8j5v7
1IF53ai7Xc0U56Y
3PZuoBAID5Wc2HD
7JYdWkrLSPkdwr4
adLQvlD726eNBSB
bvBOhCH3iADSZry
iCRJl6heRkivqQ3
ih0vzX44oOqAx2f
McdE0feGgRqW7Ca
pkci93gMrogZuBj
rGa61gmuvPhdLxV
sjndEbLyjtCKgGv
uHbuxQJl8lW7ozc
VHMLBKoKglrUVDU
wCURE6d3bPkepu2
WRmjgnKYAwPKWDb
YxYtjZvoooNbGkE
z9Y9gH1T5YWrNuG
zBIq5rxdHJRwDNY
ZnxXDlPa8U1GXgE
ZoEaEvLYb1n2sOq
zVJPv84UY57bAof
Source Key

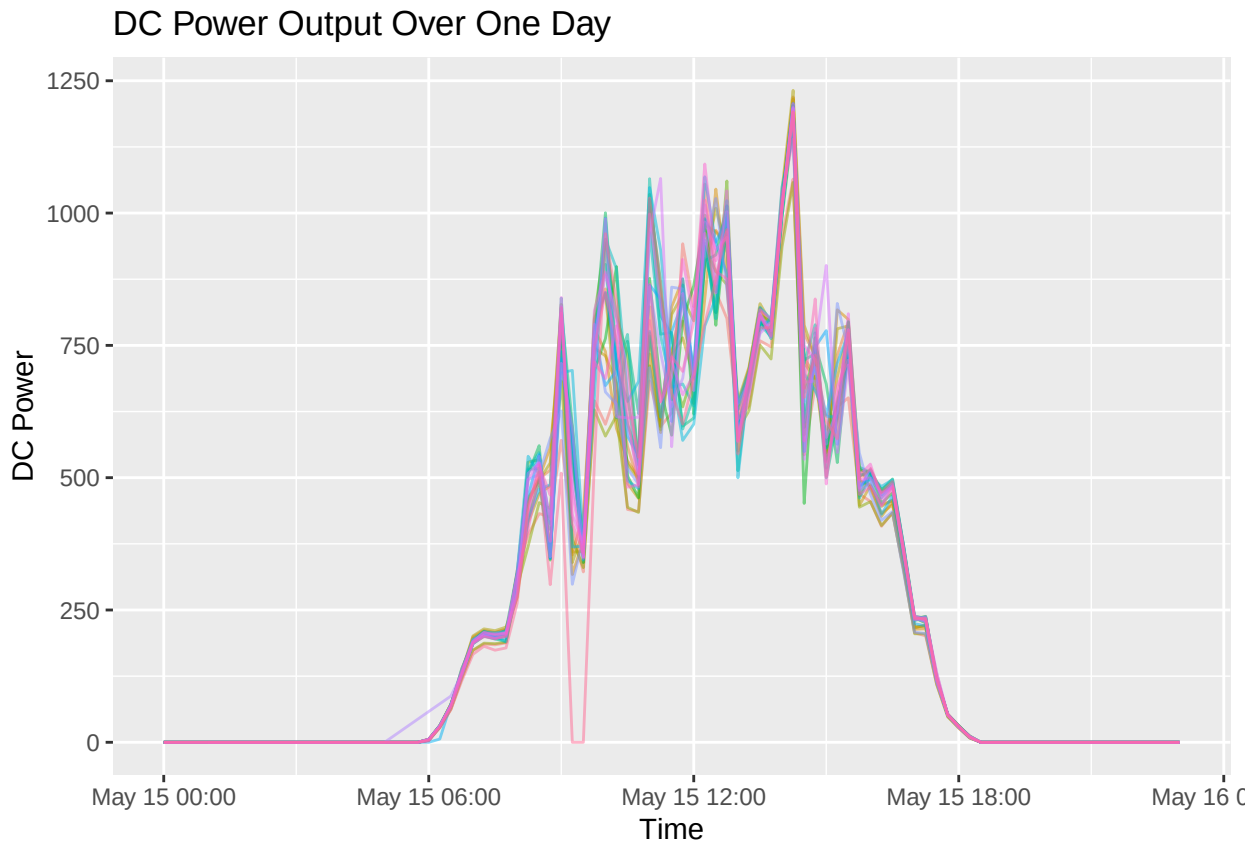


The above table and box plots show that the power generation across sources was very consistent, with a few sources showing slight variations.


Total Yield by Source
Total Yield
7500000
7000000
6500000
1BY6WEcLGh8j5v7
1IF53ai7Xc0U56Y
3PZuoBAID5Wc2HD
7JYdWkrLSPkdwr4
adLQvlD726eNBSB
bvBOhCH3iADSZry
iCRJl6heRkivqQ3
ih0vzX44oOqAx2f
McdE0feGgRqW7Ca
pkci93gMrogZuBj
rGa61gmuvPhdLxV
sjndEbLyjtCKgGv
uHbuxQJl8lW7ozc
VHMLBKoKgIrUVDU
wCURE6d3bPkepu2
WRmjgnKYAwPKWDb
YxYtjZvoooNbGkE
z9Y9gH1T5YWrNuG
zBlq5rxdHJRwDNY
ZnxXDlPa8U1GXgE
ZoEaEvLYb1n2sOq
zVJPv84UY57bAof
Source Key



Looking at the total yield observed by each source reinforces the view that each power generator at plant 435001 performed similarly. Note that the second graph accounts for the cumulative yield at the start of the month, thus showing much less variation. The first graph reveals which of the power generation sensors have been running the longest.



The above graph overlays the DC power output of each generation source during the first day. Before sunrise and after sunset, about 5:30 AM and 6:30 PM, no DC power is is generated. This is to be expected. The power output grows, peaks, and drops off. This is to be expected, if the angle of the sun is expected to impact the solar panel output. However, the data is much noisier than to be simply described by the angle of the sun.

Most sources follow similar paths, suggesting a plant-wide variable, like irradiation, is a driving force. However, we also see that one source drops to 0 output around 9:15 AM while the others only experience a dip. Sensor faults will have an impact on predictions.

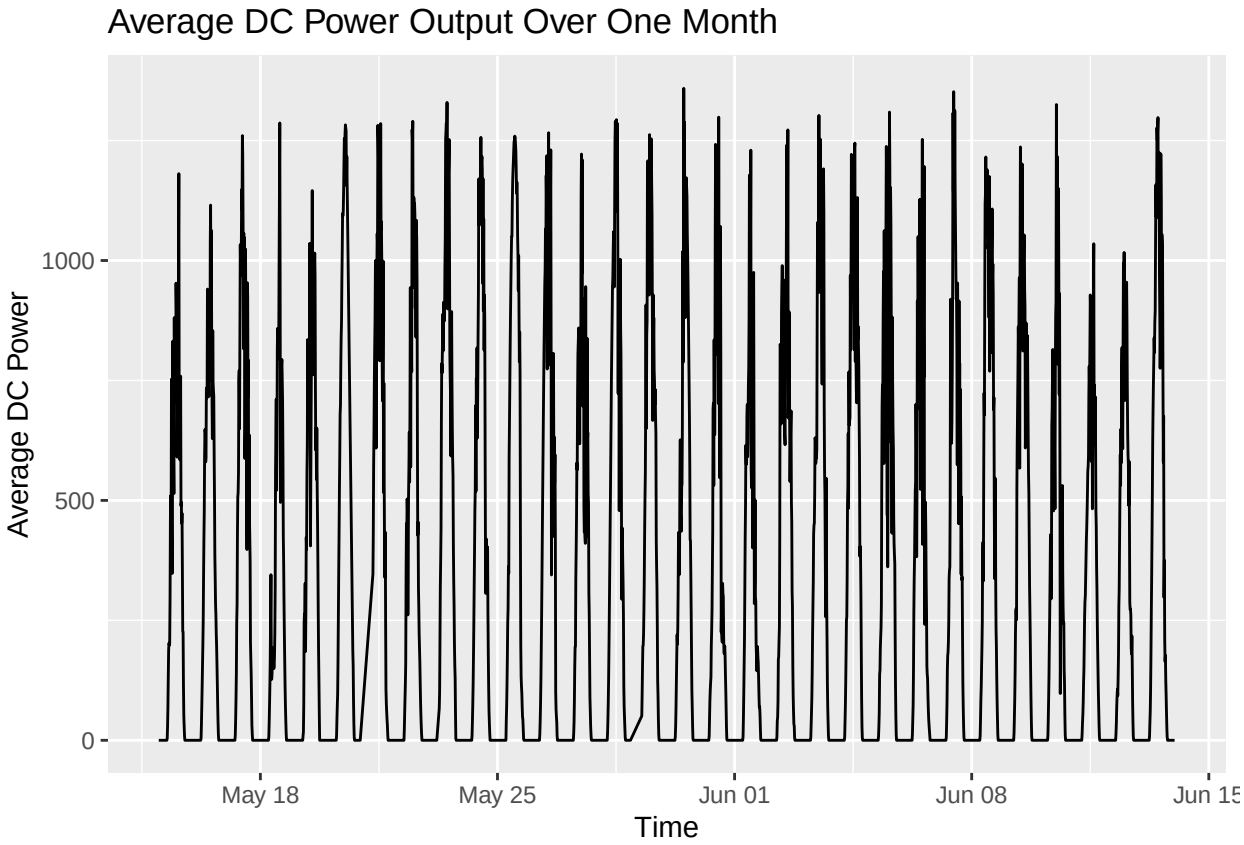
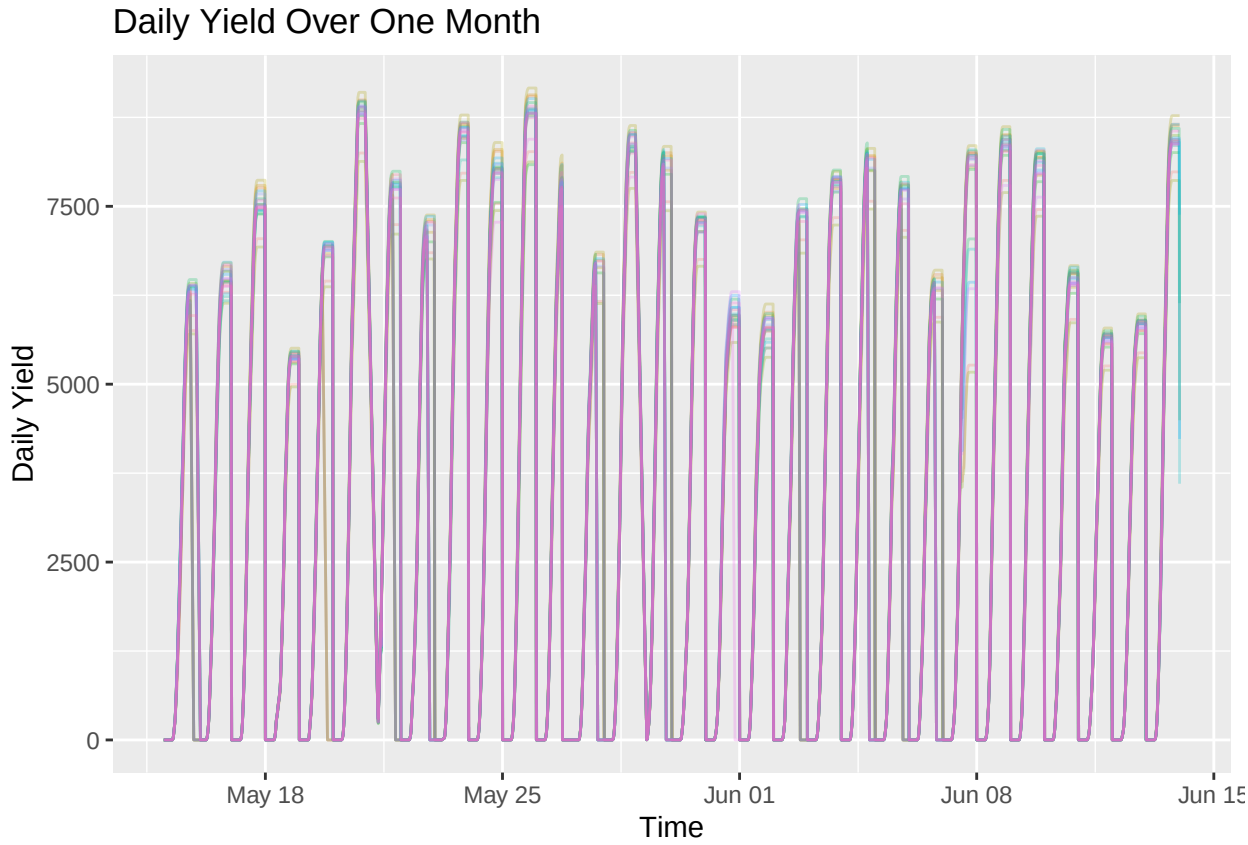

Average DC Power Output Over One Month
Average DC Power
1000
500
0
May 18
May 25
Jun 01
Jun 08
Jun 15
Time

Table 9: Summary of Weather Sensor Data

Variable	min	max	median	mean	sd
irradiation	0.0	1.15	0.026	0.233	0.306
module_temperature	18.1	65.55	24.764	31.334	12.562
ambient_temperature	20.4	35.25	24.812	25.659	3.463

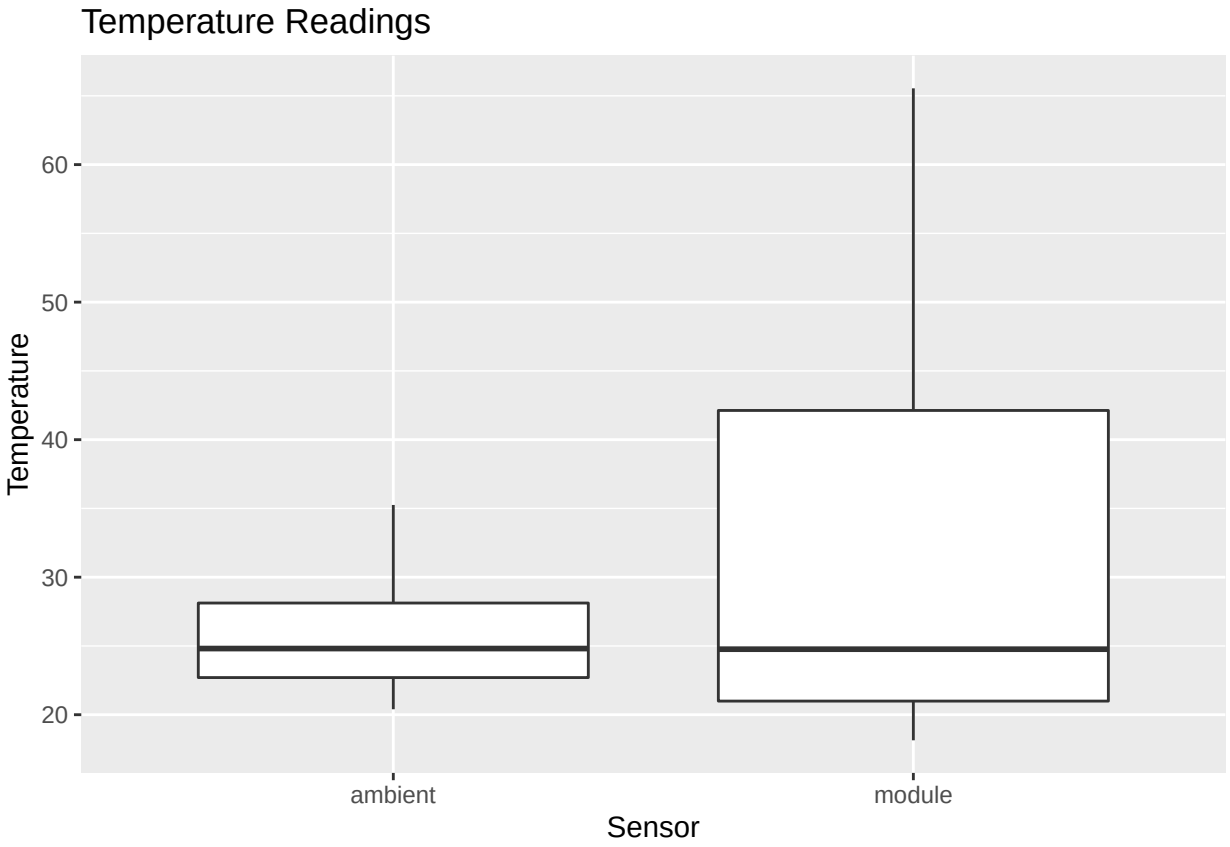


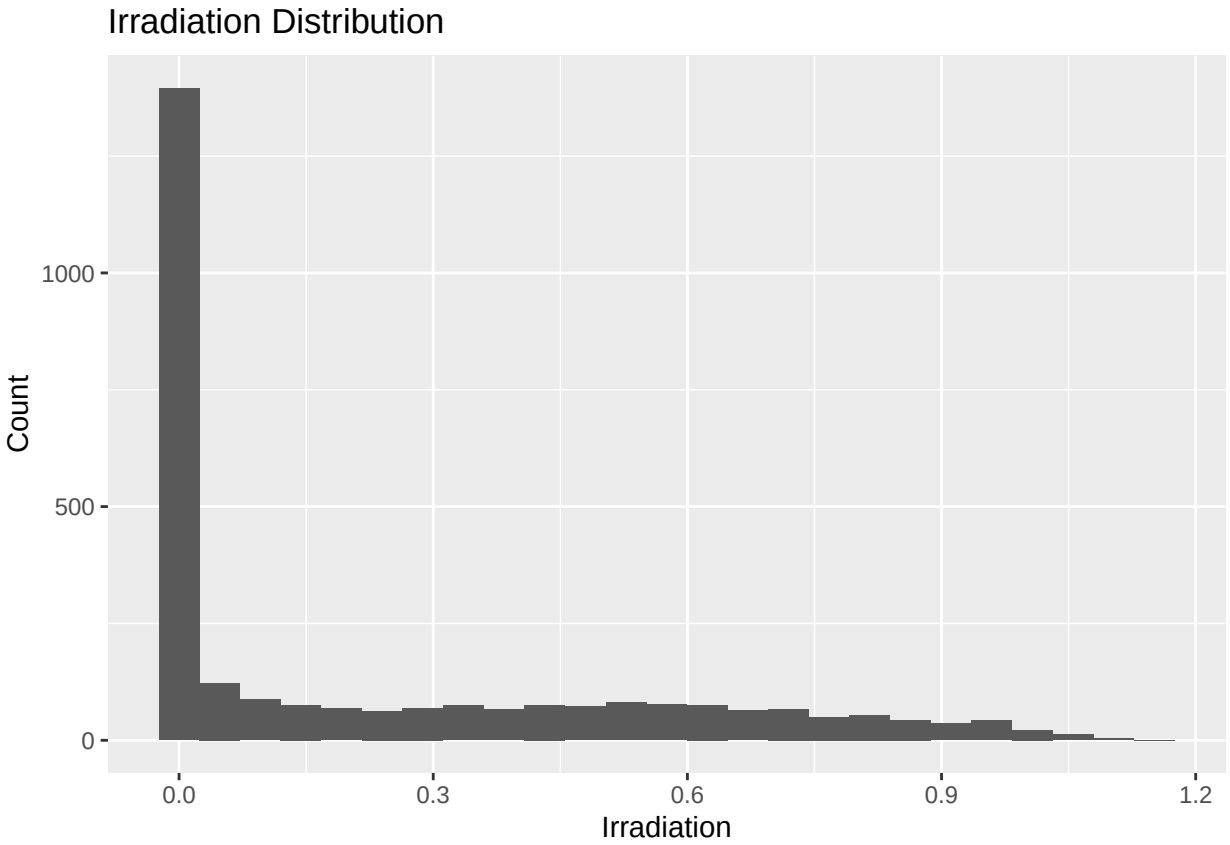
The first graph above takes the average dc output of every sensor over the month. The result is predictably periodic, and peaks at roughly the same value every day.

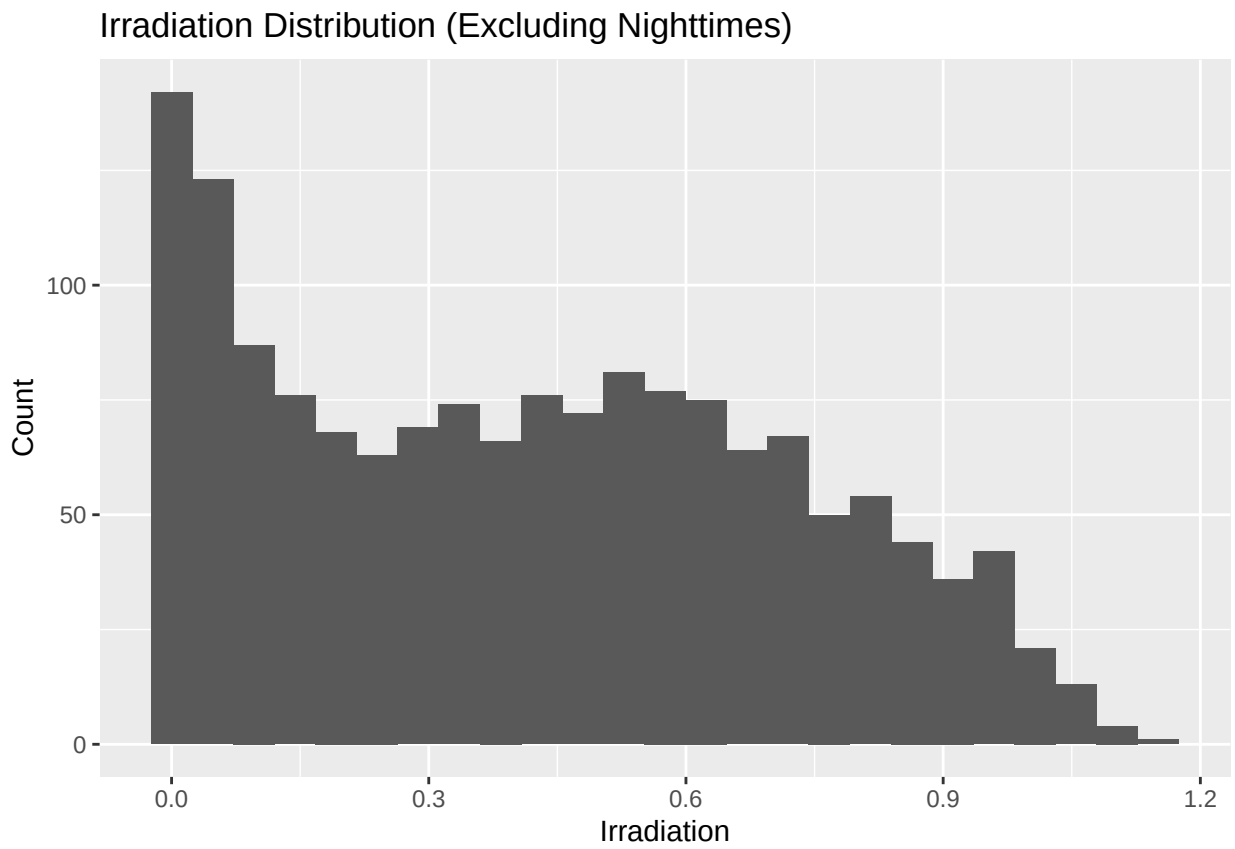
The second graph shows the cumulative daily yield of each sensor over the month. This graph more clearly demonstrates the difference in DC output from day to day, and that each solar power source outputs roughly the same amount every day.

2.4.1 Weather

It is not surprising that the module and ambient temperature have roughly the same median temperature, 24.7 degrees celsius. However, the module temperature shows a much greater spread in temperature, especially in higher directions. The module also gets colder. This suggests that the module has some sort of cooling system in an attempt to mitigate the higher temperatures that it reaches. The existence of a cooling system suggests that the module has an ideal operating temperature at which it most efficiently generates power. The box plot below visualizes the greater spread.

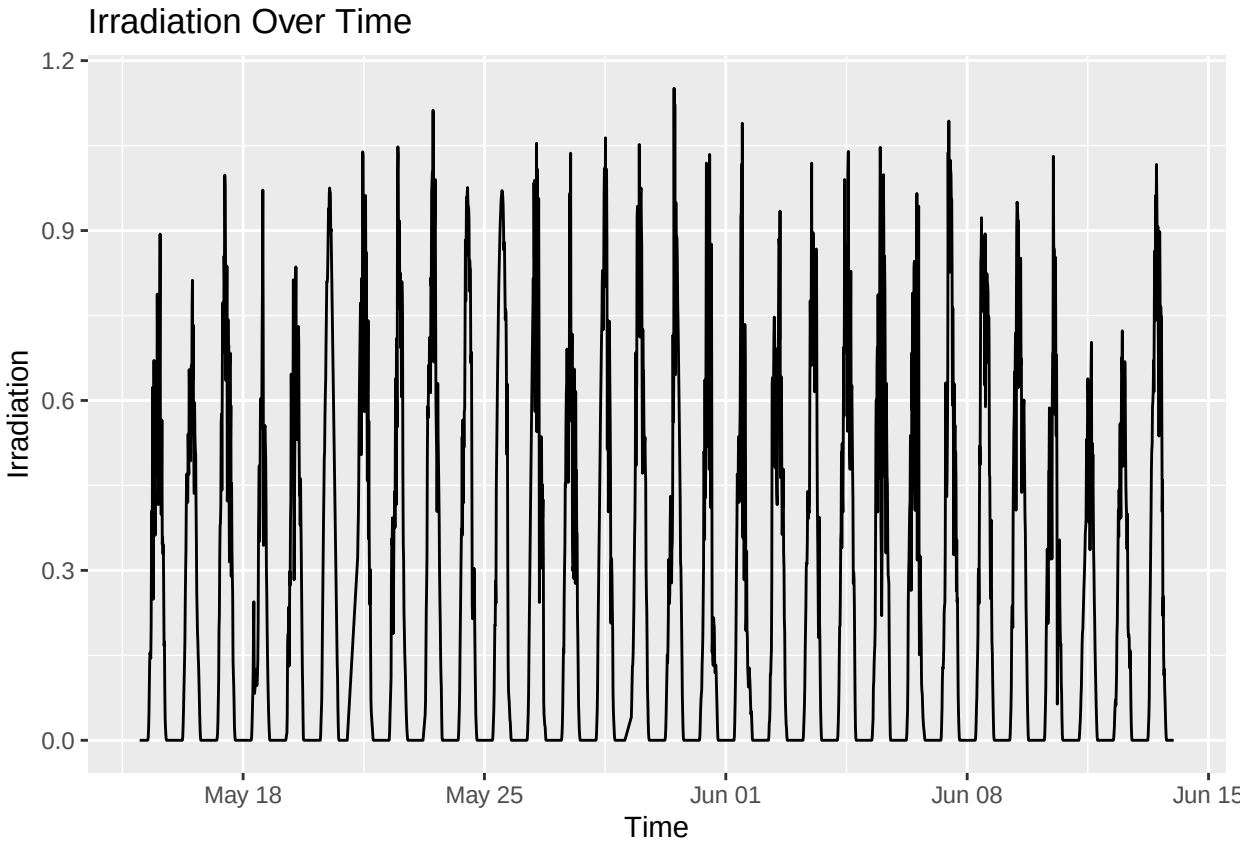

Temperature Readings
Temperature
60
50
40
30
20
ambient
module
Sensor

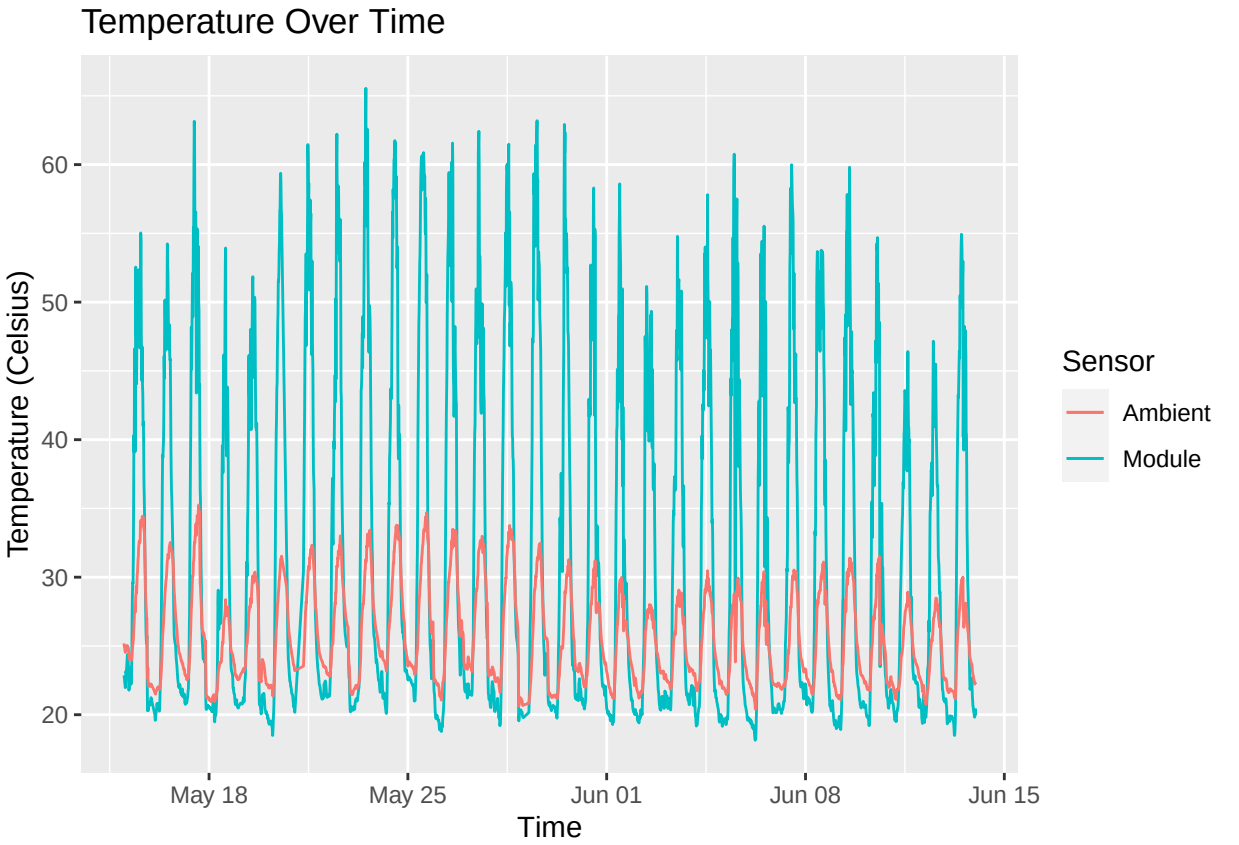

Irradiation Distribution
Count
1000
500
0
0.0
0.3
0.6
0.9
1.2
Irradiation



The first distribution above does not reveal much information. It is not surprising that the most common observed irradiation is 0, as whenever the sun is down, the measured irradiation ought to be 0.

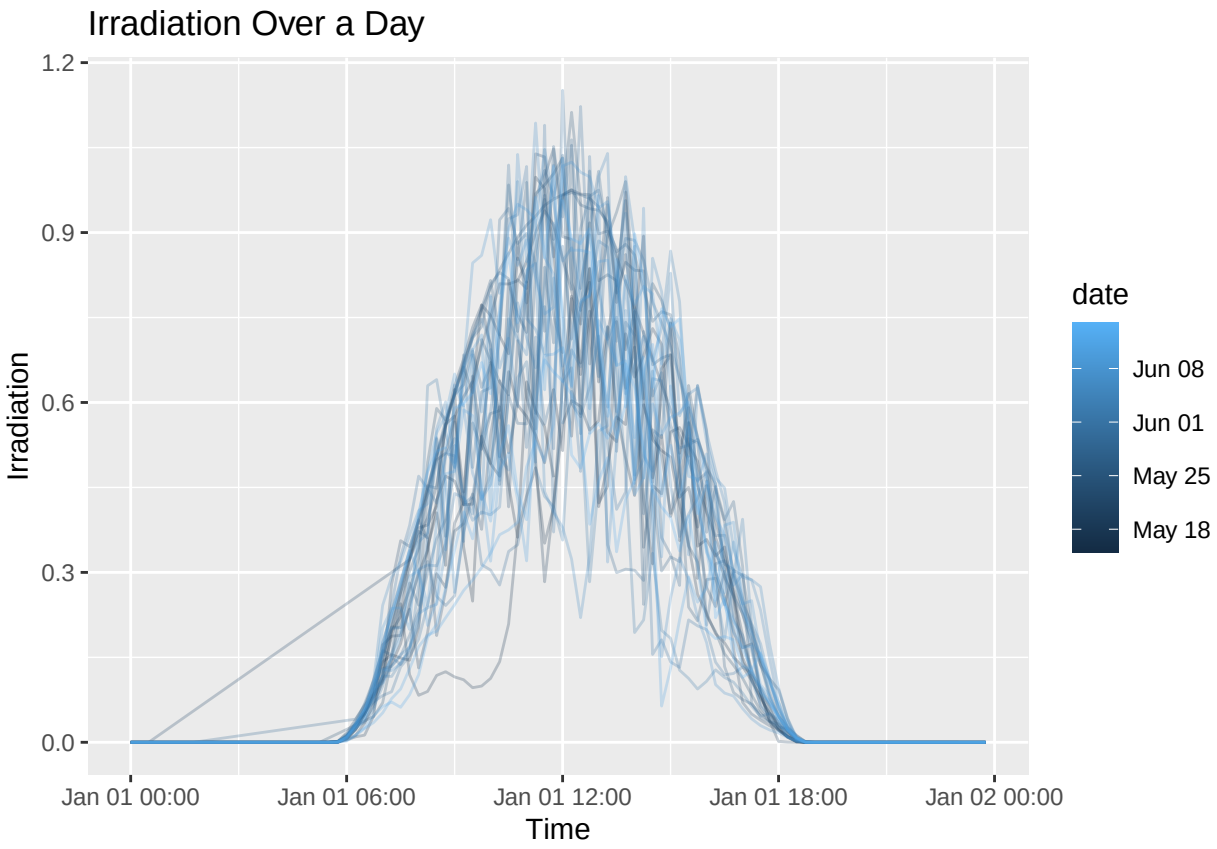
The second graph, in which nighttime observations have been removed, reveals an interesting quirk. There is a secondary mode for irradiation at about 0.5. This may have to do with the atmosphere sunlight passes through before reaching the solar panels. At certain angles, enough sunlight is scattered that the irradiation remains low. Once the angle of reflection has been breached, irradiation increases as the angle of the sunlight approaches normal, or 90 degrees. Regardless, this graph reveals there is more to the irradiation reaching the solar panels than simple geometry.

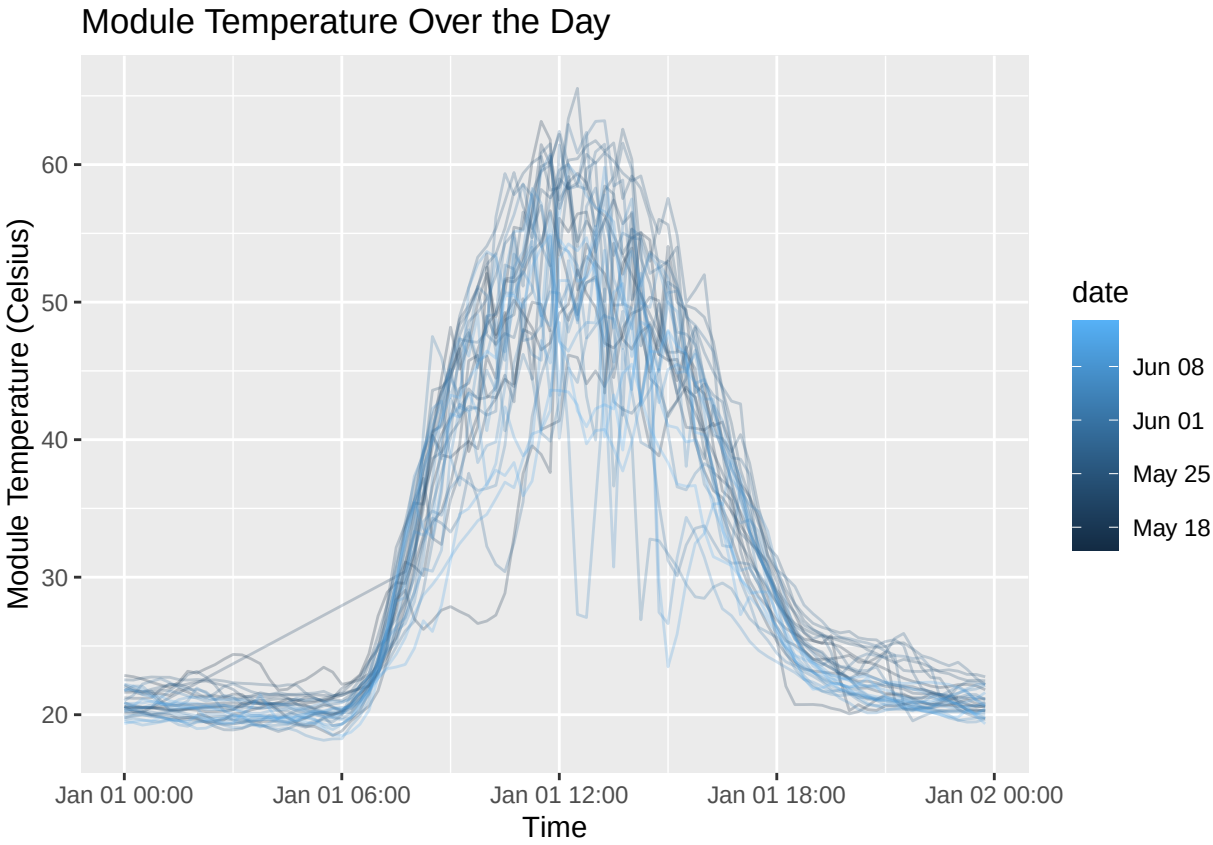

Irradiation Over Time
1.2
0.9
0.6
0.3
0.0
Irradiation
May 18
May 25
Jun 01
Jun 08
Jun 15
Time

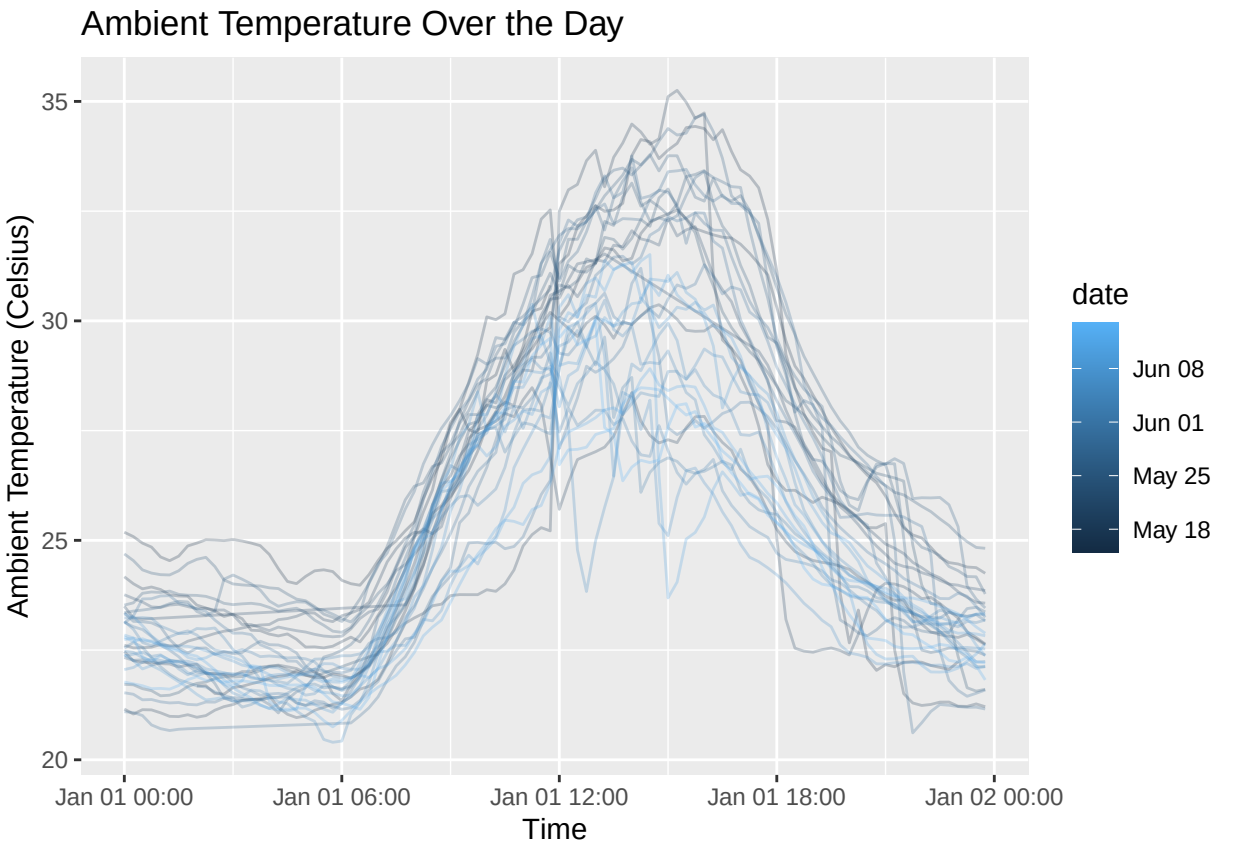


The irradiation over time follows a pattern that looks very similar to the graph generated of dc power output over time. It is periodic, with a frequency of 1 day.

The temperature data also follows what might be expected after seeing the summary table. The module temperature varies much more than the ambient temperature. However, note that the ambient temperature actually lags behind the module temperature. The relationship between irradiation and temperatures will be further examined below.


Irradiation Over a Day
1.2
0.9
0.6
0.3
0.0
Irradiation
Jan 01 00:00
Jan 01 06:00
Jan 01 12:00
Jan 01 18:00
Jan 02 00:00
Time
date
Jun 08
Jun 01
May 25
May 18

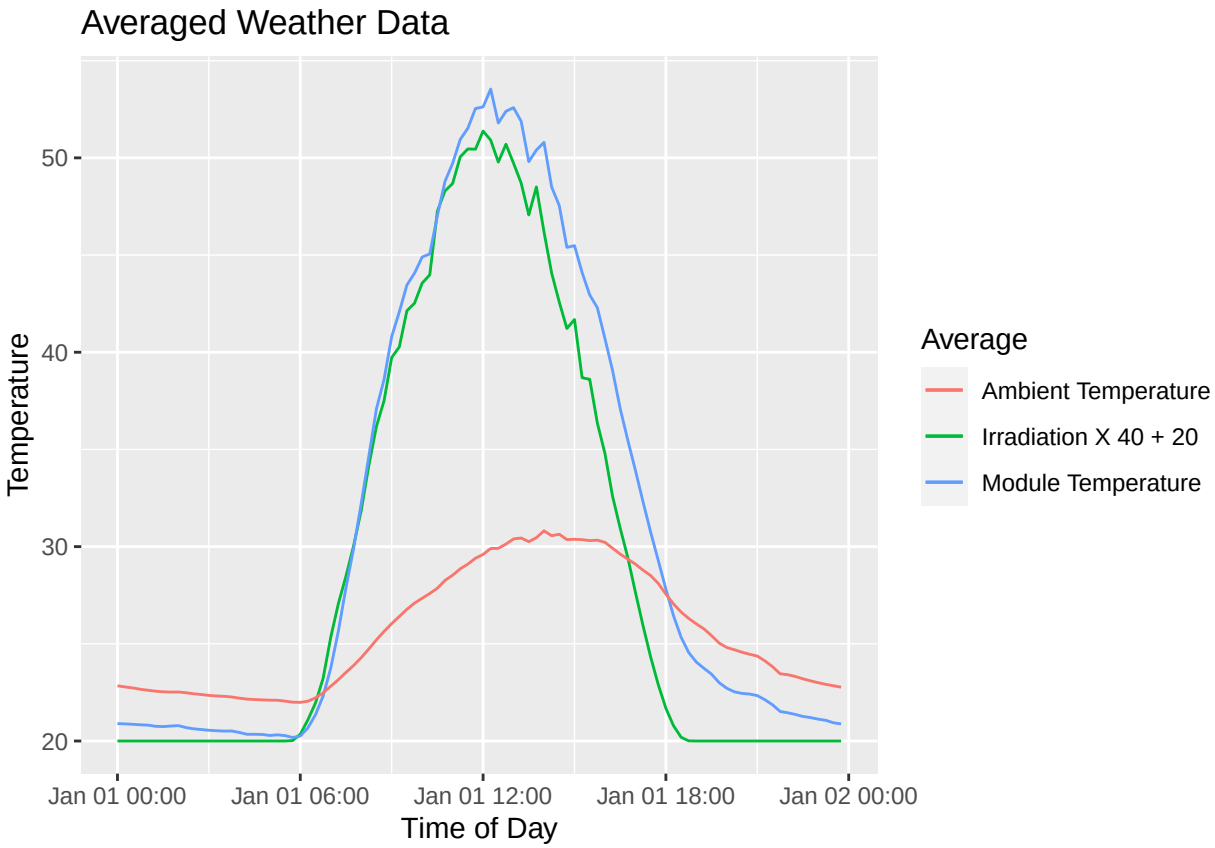

Module Temperature Over the Day
Module Temperature (Celsius)
60
50
40
30
20
Jan 01 00:00
Jan 01 06:00
Jan 01 12:00
Jan 01 18:00
Jan 02 00:00
Time
date
Jun 08
Jun 01
May 25
May 18



Irradiation appears to be very noisy. Yet it is enveloped by a parabola that represents the maximum light from the sun reaching the solar panel. The module temperature Ambient: Trends DOWN from may to june

Table 10: AC Power per DC Power

term	estimate	std.error	statistic	p.value
(Intercept)	0.277	0.006	50	0
dc_power	0.977	0.000	91637	0



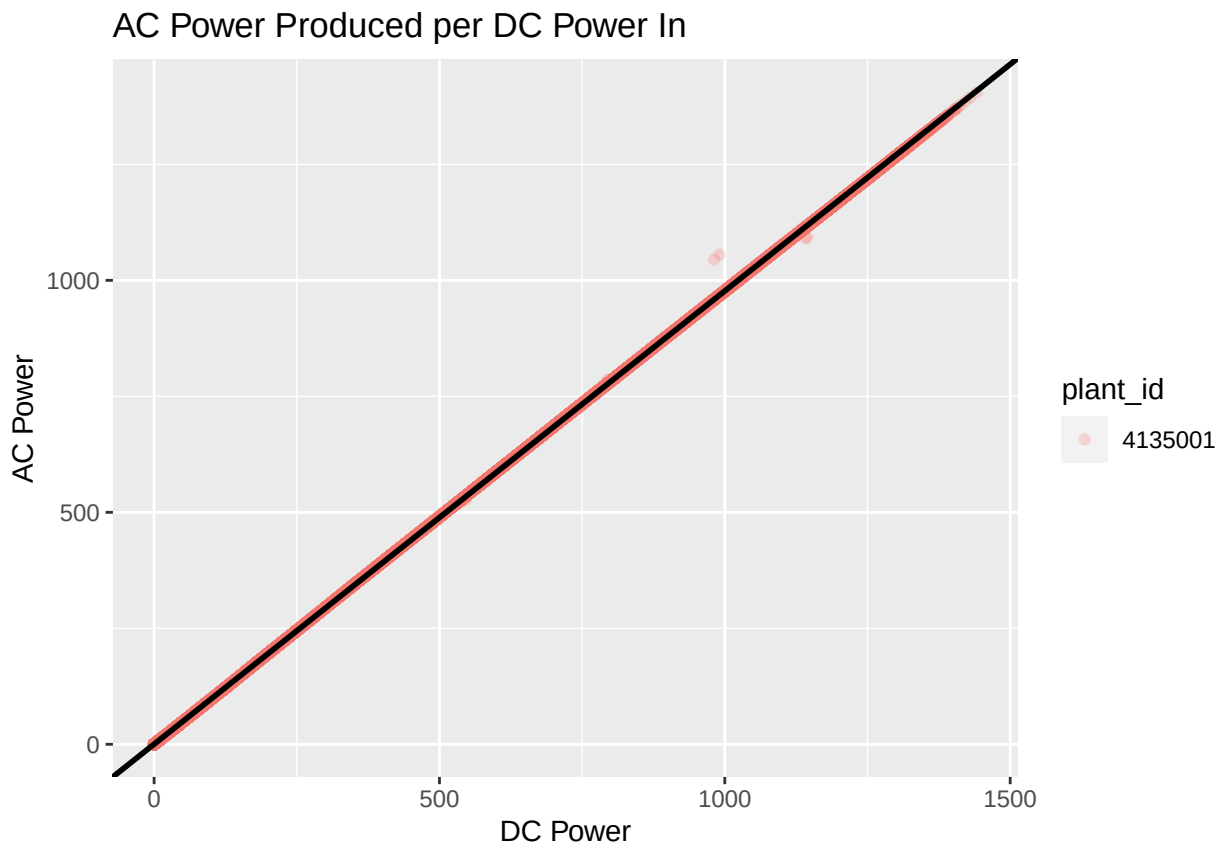
Note that in the graph above, the scale of the irradiation has been multiplied by 40 and shifted up 20 degrees to better visualize relationships. Irradiation leads module temperature by about 15 minutes and ambient temperature by a number of hours. However, the module temperature appears to be some composition of the ambient temperature and irradiation. This might suggest that the weather data can largely be described by the irradiation, or by the irradiation and the ambient temperature.

2.5 Correlation

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

The relationship between the DC Power and AC Power produced are highly correlated.

This aligns with expectations, that a certain percent of power is lost in during conversion.



This fit suggests a 97.7% conversion rate from DC to AC power, only 2.3% is lost. This suggests that the plants are using high quality transformers.

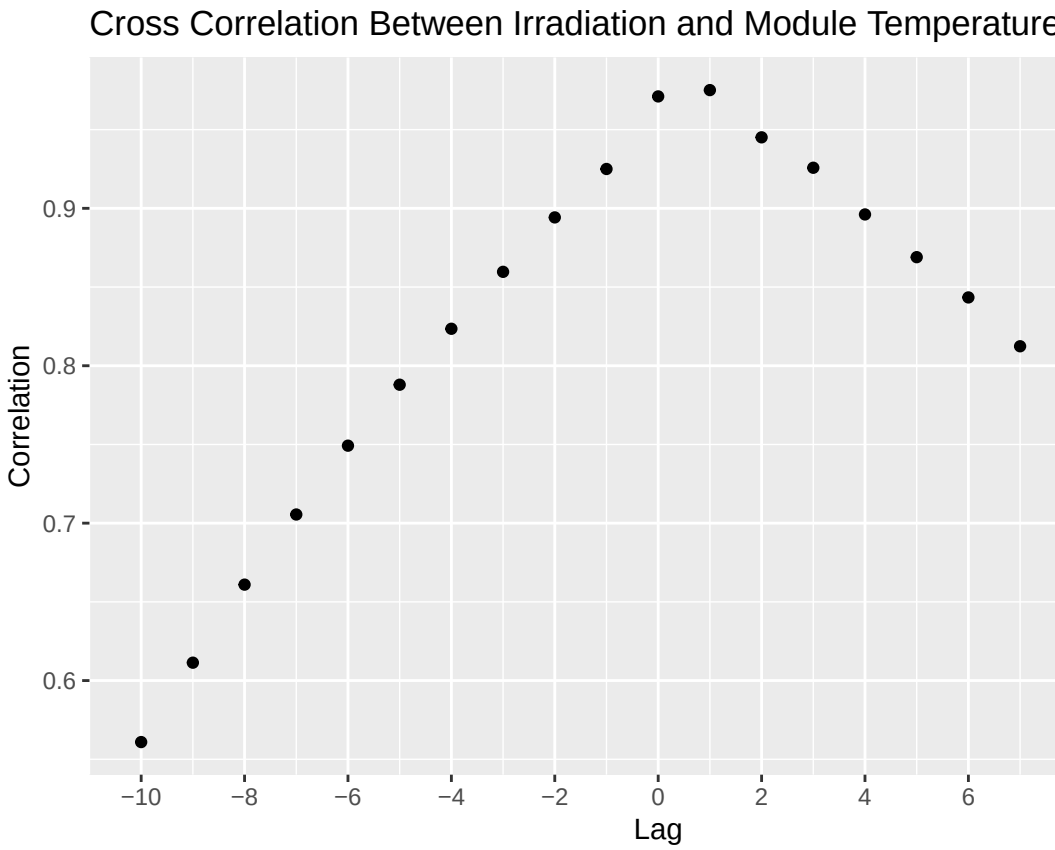
2.5.1 Lag

While exploring the weather data, a lag was noted between irradiation, module temperature, and ambient temperature. This could represent some relationship between the three. Because of the physical properties of temperature and heat flow, it is likely that irradiation drives both the module temperature and the ambient temperature. It is unlikely that the module temperature has an effect on the ambient temperature. It is possible that the ambient temperature acts as a heat or cold sink for the module.

For a better understanding of what may be going on with weather at the power plant, the values of the weather

Table 11: Lag of Maximum Cross Correlation Between Irradiation and Module Temperature

lag	correlation
15m	0.975

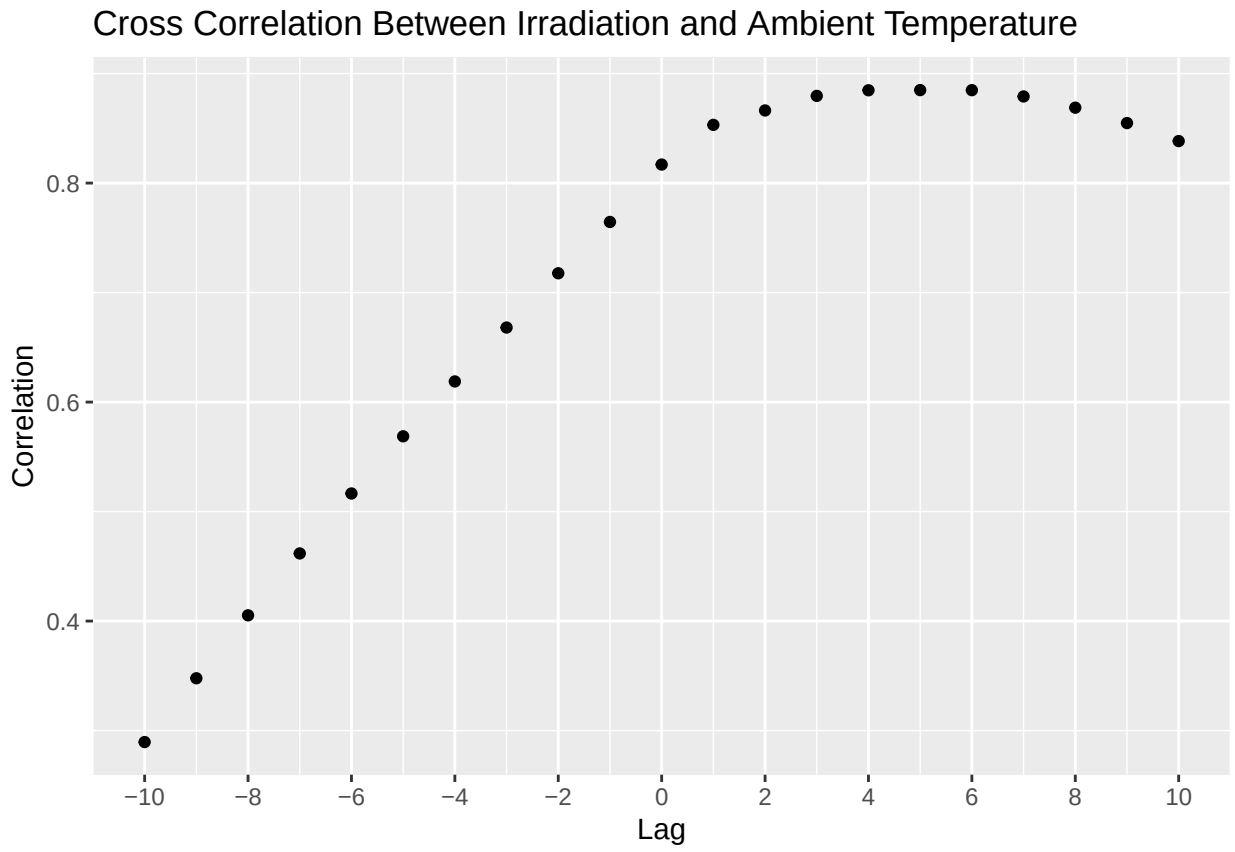


sensors are treated as time series.

A strong correlation is found between the irradiation and the module temperature after a 15 minute lag. This can be explained by the radiation heat transfer as a result of the electromagnetic radiation of the sun. It does not take long for higher irradiation to increase the temperature of the module. However, referring back to the average day temperatures graph, it appears that as the radiation increases, temperature increases in step. As the radiation decreases, the temperature decreases after about 30 minutes.

Table 12: Lag of Maximum Cross Correlation Between Irradiation and Ambient Temperature

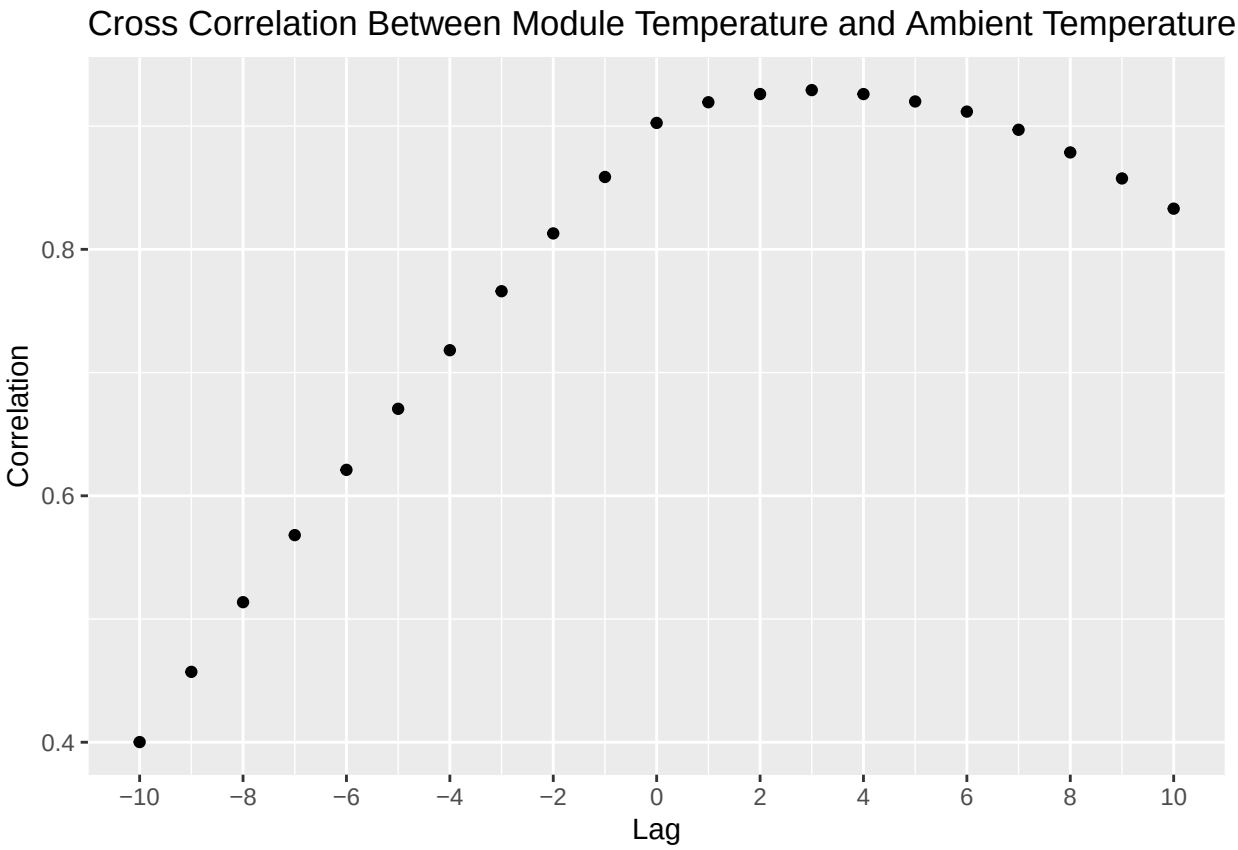
lag	correlation
75m	0.885



The connection between irradiation and ambient temperature is less clear. The correlation of 0.877 after an hour and fifteen minutes suggested that irradiation has some effect on the ambient temperature. The increased lag time has to do with how much less interaction there is between rays from the sun and air molecules vs a stationary module. However, there are sufficient confounding and unobserved variables related to weather and weather prediction that it is beyond the scope of this analysis to dig deeper.

Table 13: Lag of Maximum Cross Correlation Between Module Temperature and Ambient Temperature

lag	correlation
45m	0.929



Interestingly, the ambient temperature lags the module temperature. At first glance, this does not make much sense. It is unlikely that a module can be driving the ambient temperature.

However, if one keeps in mind the confounding variable of weather, the relationship between irradiation and module temperature explains this. It is far more likely that the ambient environment acts as a heat sink when the module becomes very hot. Though the physical interactions of heat transfer are modellable, they are outside the scope of this analysis.

2.6 The Model

2.6.1 Splitting the Train and Test Sets

A validation data set has already been separated from the data. However, to evaluate the efficacy of the model as it is constructed, it becomes necessary to further partition the data to avoid over training. The final two days of the solar data will be put into a test set, all other data remain in a train set.

2.6.2 Baseline (Naive RMSE)

As established, the RMSE is the metric by which success is measured. To get an understanding of how much the model is improving, the naive RMSE will be calculated as a baseline That is, the predicted dc_power will simply be the mean of all dc_power observed.

method	RMSE	improvement
Average by Source	401	0.003

```
## [1] 321
```

```
## [1] 402
```

The average of the dc_power over the entire data set is 321. Predicting the average dc_power results in an RMSE of 402. That is substantial. Time to explore ways to reducing reduce it.

2.6.3 Generation Source Effect

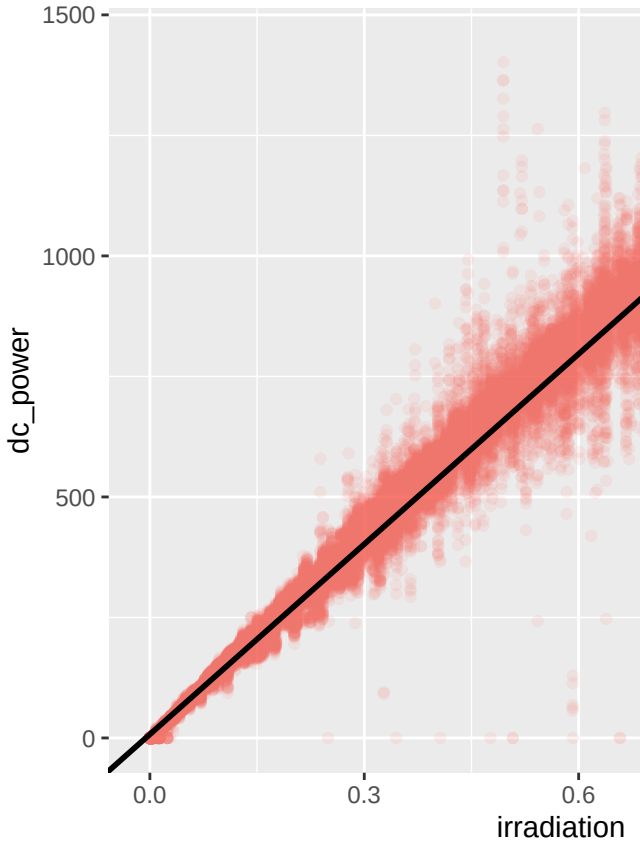
To start, the effect of stratifying by generation source is examined.

A slight improvement, but only slight. About a quarter of a percent.

2.6.4 Irradiation Effect

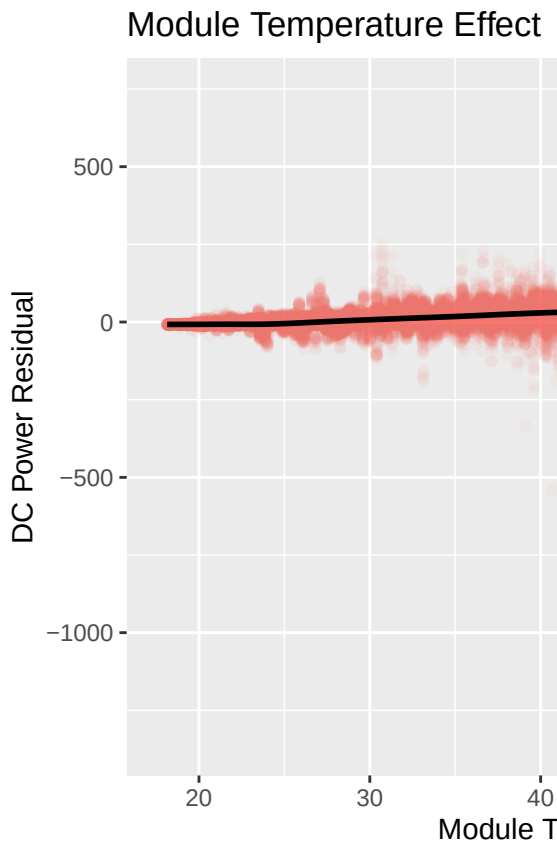
In exploration, it was observed that irradiation and dc output were highly correlated.

```
## [1] 44.4
```



44.4 is, unsurprisingly, a substatial improvement over the naive RMSE.
The above graph suggests a linear relationship fits rather well. How much can the RMSE be improved upon using a linear model, using only irradiation as a predictor?

2.6.5 Module Temperature Effect



It is not unreasonable to consider other weather effects to construct the model.
The above graph, relating module temperature and the DC power output, suggests that the module has an optimal operating temperature between 20 and 54 degrees. The relationship does not appear to be a straight line, however. Perhaps the module temperature effect is best represented by parabola

```
## [1] 40
```

40.0, a %9.9 improvement over a purely irradiation effect.

Earlier, it was discovered that module temperature lagged irradiation by 15 minutes and was highly correlated. Can a similar RMSE be achieved by simply lagging the irradiation by 15 mintutes? Note, the lagged time series results in an NA. Since this is at night, it will be replaced with the lowest predicted dc_power.

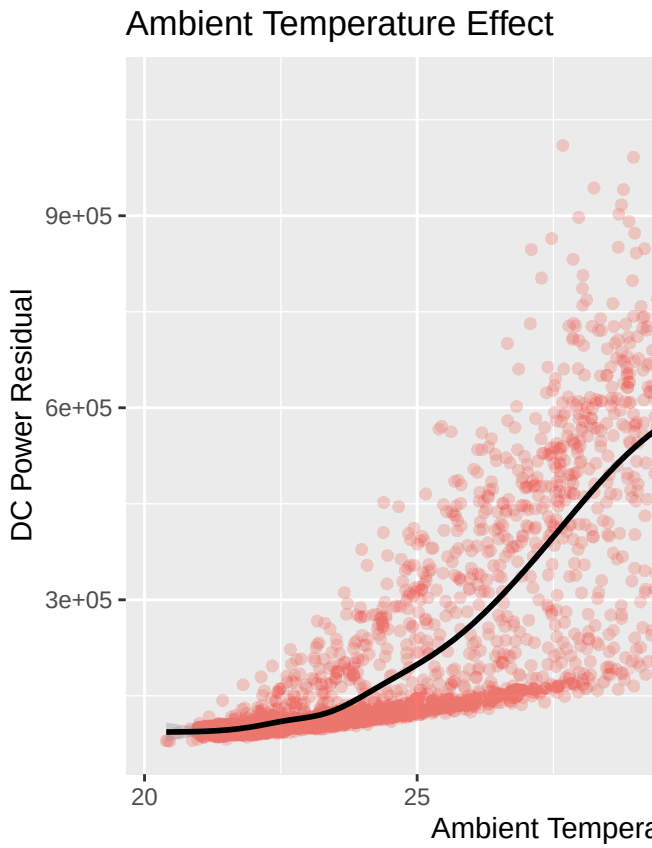
```
## [1] 40.6
```

40.61, about %1 worse than if module temperature was incorporated.

Table 14: Summary of Linear Fit Predictions

DC_Power	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Predicted	-15.1	-3.81	34.5	296	619	1316
Test Set	0.0	0.00	35.3	305	632	1374

2.6.6 Ambient Temperature Effect



Now, the residual DC power is compared to the ambient temperature.
This appears to show a negligible effect.

An improvement of less than half of a percent.

2.6.7 Clamping and Sensor Faults

Let's take a glance look at how our predictions compare to the test data.

Immediately, a source of improvement presents itself. It is not possible for negative DC power to be produced.

```
## [1] 39.7
```

Other intuitive improvements may be made. For example, consider the anomalous data found during exploratory analysis. Do we improve the predictive model if we do not train anamolous data?

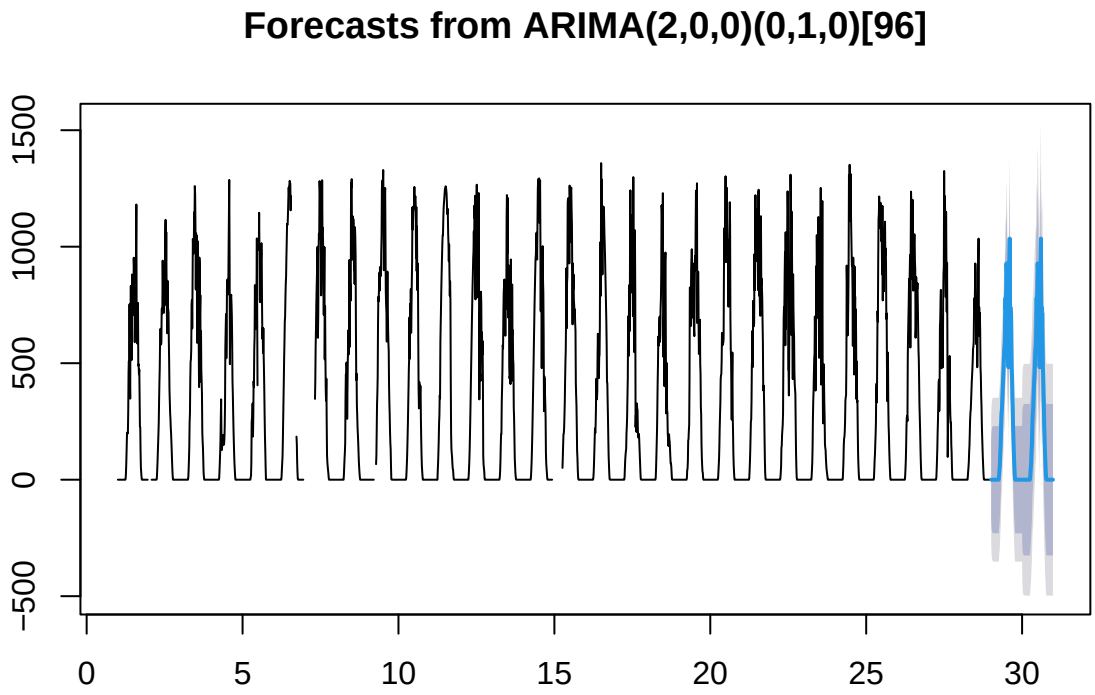
```
## [1] 39.2
```

A final RMSE of 39.15 is found with a model based on linear regression and some cleaning. 39.15 is an improvement of 90.3% over the naive RMSE, a considerable improvement.

2.6.8 ARIMA

ARIMA stands for Auto-Regressive Integrated Moving Average. It is a statistical analysis model that uses time series data to predict trends. Autoregressive: Predicts future values based on past values, regresses on its own lagged values Integrated: Differences raw observations to make the time series stationary Moving Average: Smooths out time series data by creating a subset of recent data to use as an average auto.arima (from forecast) and ARIMA (from fable) automatically determine the p, q, and d parameters, that is, the parameters used in the autoregression, integration, and the moving average.

```
## Series: .
## ARIMA(2,0,0)(0,1,0)[96]
##
## Coefficients:
##         ar1    ar2
##       0.57  0.081
## s.e.  0.02  0.020
##
## sigma^2 = 19702:  log likelihood = -15874
## AIC=31754   AICc=31754   BIC=31771
```



2.6.9 ARIMA By Generation Source

2.6.10 Random Forest

2.6.11 PCA

Principle Component Analysis

method	RMSE	improvement
Average	401.8	0.000
Average by Source	400.6	0.003
Irradiation Effect	44.4	0.890
Irradiation + Module Temperature Effect	40.0	0.900
Irradiation + Irradiation Lag Effect	40.6	0.899
Irradiation + Module + Ambient Temperature Effect	39.9	0.901
Clamped	39.7	0.901
Clamped + Anomolies Removed	39.2	0.903
ARIMA	179.1	NA
ARIMA by Source	41.0	NA

3 Results

3.1 Model Root Mean Squared Errors

3.2 Final Validation

4 Conclusion