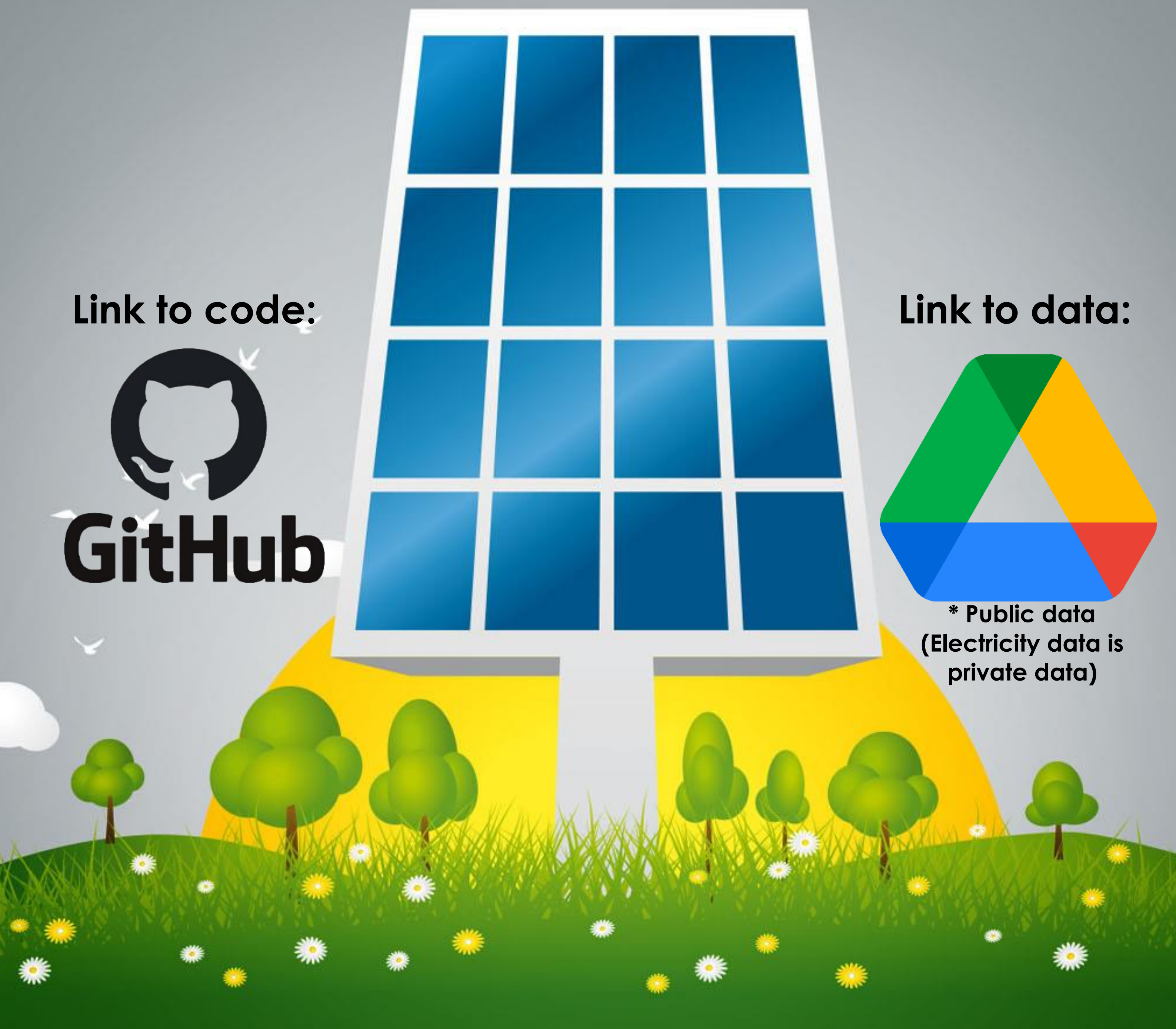
ADVANCED DATA ANALYSIS IN R

Solar Panels

Yarin Shohat, Avi Elbaz, Tomer Ronen
And Maxim Lisiansky



Link to code:

GitHub

Link to data:

* Public data
(Electricity data is
private data)

1. Introduction:

This research focuses on one of the most exciting and discussed fields in recent years: Renewable Energy.Our research question concerns the production of electricity using solar energy: **"How can the optimization of solar panel placement influence electricity productivity in different geographic locations and climatic conditions?"** The general problem area to which this analysis contributes is the optimization of renewable energy systems, the specific problem that we trying to solve is not optimal placement for solar panels specifically in Israel.

As we know, electricity production with solar panels is made possible by the sun's radiation and is affected when the radiation is obstructed. Since the discovery of this field, global development has demonstrated that climate conditions in different locations on Earth can lead to variations in electricity production using solar panels. The optimal location for solar panels varies by country, and while many countries have made significant advancements in this field, Israel has lagged behind.

In Israel, the electricity sector is monopolized by the Electric Company. Although the electricity revolution in Israel at the past years has opened up the possibility of private sector production, the measurement data (for electricity production across various locations) remains confidential due to commercial reasons. Therefore, prior work in this field is limited in Israel, making it hard to extend this knowledge to all in our country.

The importance of this research is to support the evolving field of solar energy in Israel by providing public access to information on the effectiveness of solar panels in different locations and the impact of climatic conditions on their efficiency. This will help to innovate and transparent the renewable energy sector in Israel. Optimizing of solar panel placement can significantly enhance the efficiency and productivity of solar energy systems, contributing to Israel's renewable energy goals and reducing reliance on non-renewable energy sources.

Our approach is to aggregate the solar and weather company data sets, by using regression models to predict the electricity productivity based on the meteorological data of the solar farms. The difficult here is to find a company which willing provide their data. Additionally, our understanding about weather data was difficult since we lack expertise in geophysics and have limited knowledge of R and the methods we can use.

2. Data Overview:

The dataset contains observation points collected daily throughout the year 2023 from nine distinct locations in Israel. Each observation point provides statistical data for that specific day, including minimum, maximum, and average values, or other relevant labels.

We have one entity which is an observation point and its key is Location and Time Stamp. All records represent daily observation points for each location, totaling 3,281 records. Our unit of analysis is the daily electricity output, measured in kilowatts per hour and normalized by the size of the solar farm, regardless of the specific farm to which the solar panel belongs. The family of climate features were selected for analysis due to its strongest statistical significance, as determined by our model.
The dataset includes several categories of features, such as:

- **Electricity Features**: Data collected from the solar company, including variables such as Daily Energy Yield, AC (Alternating Current), and DC (Direct Current).
- **Weather Features**: Data collected from the Israel Meteorological Service and VisualCrossing, including variables such as Temperature, Rainfall, Radiation, Daylight Hours, and more.
- **Time Features**: Variables indicating the time-related aspects of the observation, such as the Month and Season.
- **Location Features**: Variables related to the geographical attributes of the observation point, such as Altitude and the general Area within Israel.

The comprehensive integration of these features allows us a percise analysis of solar energy production and its correlation with various climatic and locational factors.

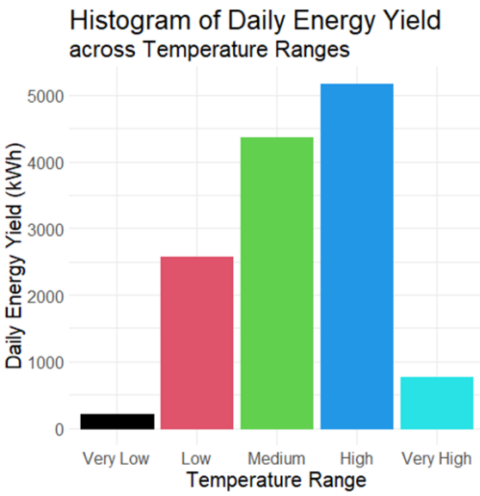
3. Methods and Results:

Preprocessing: We had a large preprocessing to do because we used real raw data. We merged the data from 5 different sources (Electric company- 9 files, Israel Meteorological Service – 9 files, Wikipedia – 1 file, VisualCrossing – 9 files and Google Maps), some of them with different time measure units (10 min interval/daily interval). Every data had different column names and we needed to merge the names into a column. We created new columns like daylight hours from the columns sunrise and sunset and created dummy variables to prepare the linear regression model.

We handled NA values with interpolation, because it's an excellent way to determine the unknown values that lie in between the known data points. We have about 365 records for every location and very small amount of those records had NA values.

Key results: We have various findings on the effects of different variables from the different families of features on energy yield, but we chose to present the following three key findings:

1. High enough temperature exhibits a low to negative correlation with Energy Yield

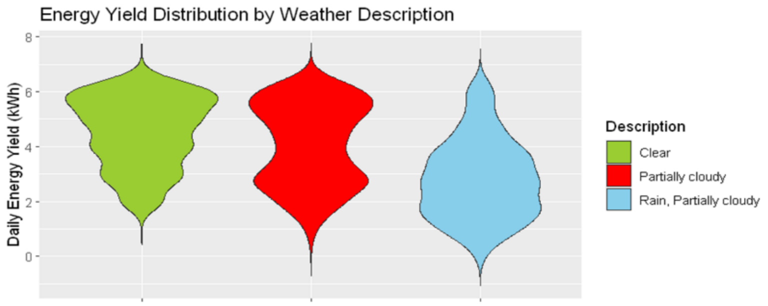


We discovered this finding by analyzing relationships and correlations through various visualizations. These revealed that while Temperatures positively correlate with energy yield, too high temperatures statistically lead to increased electricity output, excessively high temperatures can harm production, as shown in the histogram graph. When we tried to figure out what the optimal temperature is, we consulted with an electrical engineer from the solar company, who explained that each panel has a sensor that shuts it down to prevent damage from overheating. Due to the field's novelty, they haven't yet addressed the issue of panel overheating, that temperature is different for each panel.

This finding is significant and indicates a need for further investigation into the intersection of electricity production and solar energy under extreme temperatures, which may be explored in future studies due to the current project's time constraints.

This finding led us to further investigate the various climate variables, resulting in another interesting discovery:

2. The cloudiness factor has no clear statistically significant relationship with energy yield.



While we initially expected to see a clear negative correlation between cloudiness and electricity output, our analysis revealed that cloudiness does not necessarily indicate a drop in the output. Different types of clouds can still allow for near-optimal production, demonstrating that cloudiness alone is not a definitive predictor of reduced energy yield.

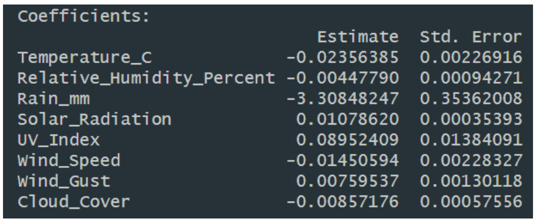
As we can see the violin plot shows us that even if it is a partially cloudy day, we can still get a fairly high daily energy yield. But we can also observe that rain has some negative impact on the energy production, which we wanted to deepen our analysis in.

3. The strongest factor affecting the output of the panel is rain.

Using a multivariable regression model with climate variables, we aligned our goal with understanding and predict how climatic conditions impact electricity generation and optimizing solar panel placement accordingly.

Given Israel's diverse climate conditions, we determined that climate-related variables offered the most relevant data for our research question, ensuring our model focuses on maximizing efficiency based on these variations. Therefore, we focused on the model that used the climate conditions variables.

During the model creation, we initially encountered poor metrics (e.g., RMSE of 1.151). Realizing the dependence of each record on the panel's position, we implemented a fixed-effect model, incorporating the panel position variable using a factor function. This adjustment significantly improved our metrics (e.g., RMSE decreased to 0.565), leading us to proceed with this model.

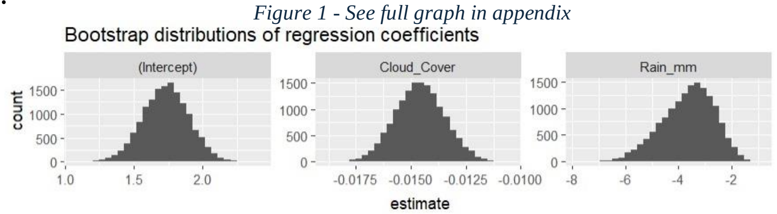


```
Coefficients:
                            Estimate  Std. Error
Temperature_C            -0.02356385  0.00226916
Relative_Humidity_Percent -0.00447790 0.00094271
Rain_mm                  -3.30848247  0.35362008
Solar_Radiation           0.01078620  0.00035393
UV_Index                  0.08952409  0.01384091
Wind_Speed               -0.01450594  0.00228327
Wind_Gust                 0.00759537  0.00130118
Cloud_Cover              -0.00857176  0.00057556
```

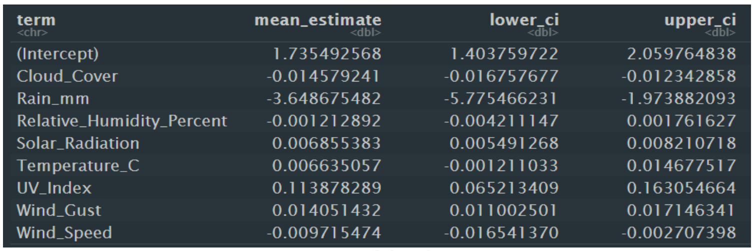
Fixed Model Regression - Coeff Table

The regression findings are as follows:
After exploring various models, we discovered that rain has a very strong negative effect on electricity output, more so than radiation, clouds, and temperature. Therefore, when determining the optimal geographic location for solar panels, the rain variable may carry the most weight.

Figure 1 - See full graph in appendix

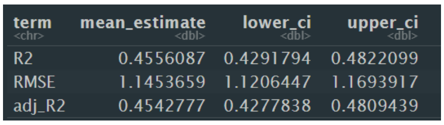


To ensure the model results are well-supported by the data, we used Bootstrap because of the limited number of observations and to further deepen our understanding and weight of each co-efficient on the model, we decided to use bootstrap. By using bootstrap, we generated a lot more observations and created a much-fitted model. As the co-effs table shows, the results we got is like those from the linear regression model, thus confirming our analysis about the rain variable, which is that the rain variable does indeed have the most impact on the model. We can also see a visualization of this finding in the rain graph (rightmost) which is skewed to the right.



term <chr>	mean_estimate <dbl>	lower_ci <dbl>	upper_ci <dbl>
(Intercept)	1.735492568	1.403759722	2.059764838
Cloud_Cover	-0.014579241	-0.016757677	-0.012342858
Rain_mm	-3.648675482	-5.775466231	-1.973882093
Relative_Humidity_Percent	-0.001212892	-0.004211147	0.001761627
Solar_Radiation	0.006855383	0.005491268	0.008210718
Temperature_C	0.006635057	-0.001211033	0.014677517
UV_Index	0.113878289	0.065213409	0.163054664
Wind_Gust	0.014051432	0.011002501	0.017146341
Wind_Speed	-0.009715474	-0.016541370	-0.002707398

Regression Model with Bootstrap



term <chr>	mean_estimate <dbl>	lower_ci <dbl>	upper_ci <dbl>
R2	0.4556087	0.4291794	0.4822099
RMSE	1.1453659	1.1206447	1.1693917
adj_R2	0.4542777	0.4277838	0.4809439

Bootstrap Model Measurements

Lastly, another model was made with coefficients regarding the location such as altitude, area and location. To include this model variables and the previous model which included variables from the weather family, we combined those two models into one using weighting. Each model was given weight W, the model who showed better results (weather model) was given increased weight. Then we changed the weight of every model empirically to see which weight distribution gives the best results.

	R²	**R².adj**	**RMSE**	**P-value**
Weather Model	0.79339	0.79237	0.5655196	< 2.22e-16
Location Model	0.53881	0.53754	0.8449048	< 2.22e-16
Combined Model	0.6359047	0.633671	0.9370887	0

We discovered that the best Combined Model is 0.6 Weather Model and 0.4 Location Model, that gives us the best results, as we can see the weather model still have better measurements, but the combined model includes more variables and offers decent results.

Further work needs to be done to improve the models.

4. Limitations and Future Work:

Our research faced several limitations: The study was constrained by the small number of locations examined, with only nine sites providing insufficient geographical diversity for a deep analysis. We also encountered difficulties with data consistency across locations, often finding different features missing from various sites forcing us to give up on them.

Given additional time, we would enhance our study by expanding the number of research locations to increase geographical representation and conduct more thorough research to acquire missing weather and geographic features for each location. Moreover, we would seek expert consultation from geophysicists or collaborate with the Department of Earth and Environmental Sciences at Ben Gurion University to conduct a more in-depth analysis of weather-related factors. Furthermore, we would investigate the practical considerations for solar panel farm construction, including regulatory guidelines, restrictions, and tolerance for extreme weather conditions.

Appendix

Bibliography

- Israel Meteorological Service - https://ims.gov.il/he
- VisualCrossing - https://www.visualcrossing.com/
- Wikipedia- Geography_of_Israel - https://en.wikipedia.org/wiki/Geography_of_Israel
- Kan 11- Hayot Kis - https://www.kan.org.il/content/kan/podcasts/p-8127/31796/
- Israel Meteorological Service - https://ims.gov.il/he

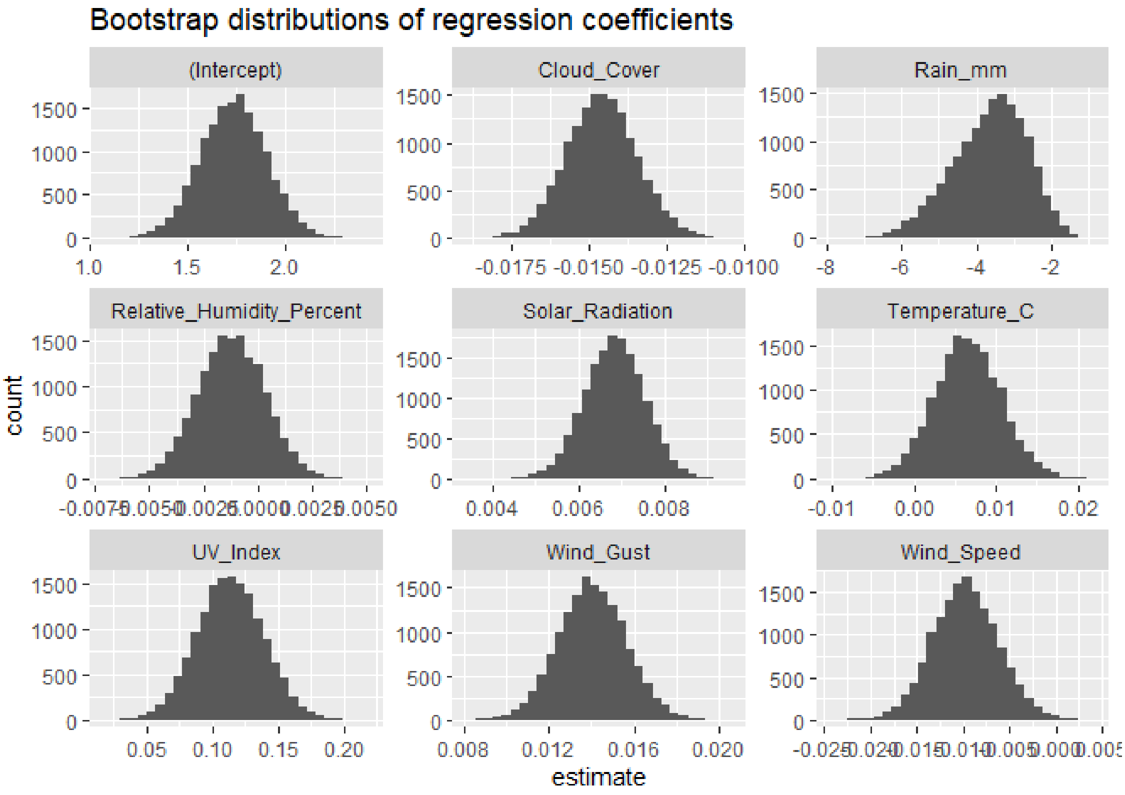
Preprocessing

More detailed explanation of our preprocessing steps and the specific variables involved, please refer to the “README” section included in our documentation, including:

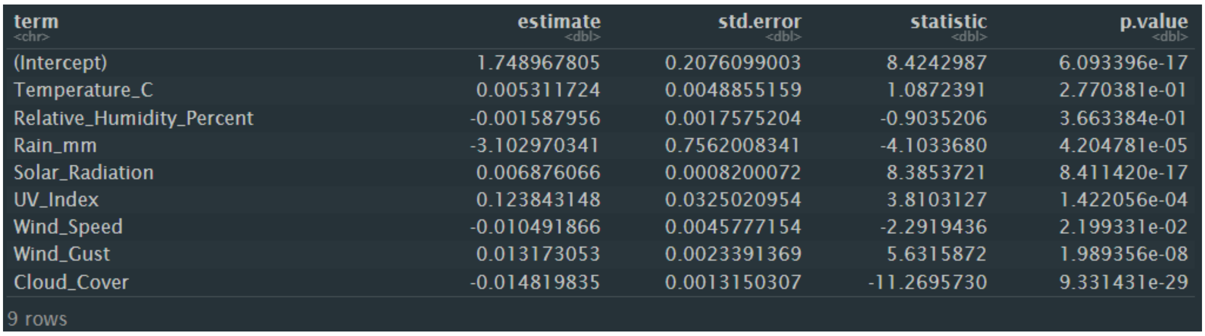
- Feature Engineering – Creating new columns from existing columns
- Dummy Variables for columns we will use in the Linear Regression Model
- Handling NA values with Interpolation
- Merge 28 files to one Data Frame

Figure adding

- Figure 1: Bootstrap distributions of regression coefficients

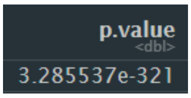


- Linear regression without Fixed Model

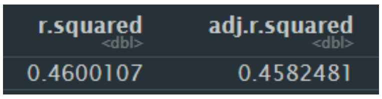


term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	1.748967805	0.2076099003	8.4242987	6.093396e-17
Temperature_C	0.005311724	0.0048855159	1.0872391	2.770381e-01
Relative_Humidity_Percent	-0.001587956	0.0017575204	-0.9035206	3.663384e-01
Rain_mm	-3.102970341	0.7562008341	-4.1033680	4.204781e-05
Solar_Radiation	0.006876066	0.0008200072	8.3853721	8.411420e-17
UV_Index	0.123843148	0.0325020954	3.8103127	1.422056e-04
Wind_Speed	-0.010491866	0.0045777154	-2.2919436	2.199331e-02
Wind_Gust	0.013173053	0.0023391369	5.6315872	1.989356e-08
Cloud_Cover	-0.014819835	0.0013150307	-11.2695730	9.331431e-29

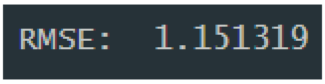
9 rows



p.value <dbl>
3.285537e-321

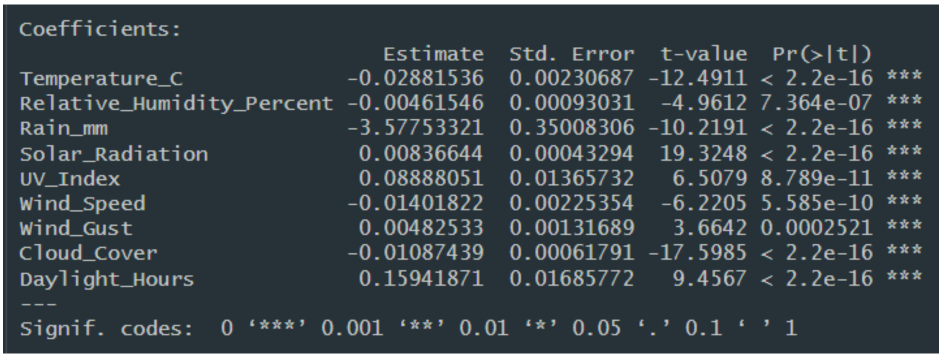


r.squared <dbl>	adj.r.squared <dbl>
0.4600107	0.4582481

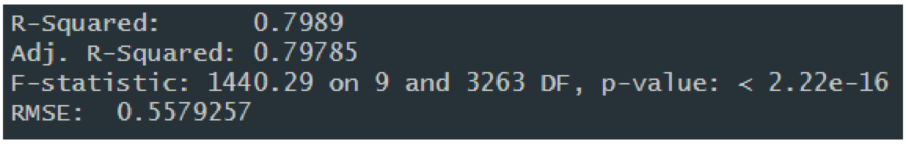


```
RMSE:   1.151319
```

- Linear regression with Fixed Model



```
Coefficients:
                            Estimate  Std. Error  t-value  Pr(>|t|)
Temperature_C             -0.02881536  0.00230687 -12.4911 < 2.2e-16 ***
Relative_Humidity_Percent -0.00461546  0.00093031  -4.9612 7.364e-07 ***
Rain_mm                   -3.57753321  0.35008306 -10.2191 < 2.2e-16 ***
Solar_Radiation            0.00836644  0.00043294  19.3248 < 2.2e-16 ***
UV_Index                   0.08888051  0.01365732   6.5079 8.789e-11 ***
Wind_Speed                -0.01401822  0.00225354  -6.2205 5.585e-10 ***
Wind_Gust                  0.00482533  0.00131689   3.6642 0.0002521 ***
Cloud_Cover               -0.01087439  0.00061791 -17.5985 < 2.2e-16 ***
Daylight_Hours             0.15941871  0.01685772   9.4567 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
R-Squared:      0.7989
Adj. R-Squared: 0.79785
F-statistic: 1440.29 on 9 and 3263 DF, p-value: < 2.22e-16
RMSE:   0.5579257
```