# Solar Radiation Prediction

Shunki Akiyama

March 2021

Contents

# 1 Introduction

## 1.1 Project Description

The purpose of this project is to analyze what factors affect the amount of solar radiation in general. The dataset used in this project is originally provided by NASA and it contains solar radiation and some weather information. It is critical for us to know how to predict solar power generation because we need to rely on clean energy more in the future. In this analysis, we will look at four months' data and figure out the correlation between variables.

## 1.2 The Data Sets

The dataset used for this project is taken from the website Kaggle.com. It is available at the below link: https://www.kaggle.com/dronio/SolarEnergy

The dataset is meteorological data from the HI-SEAS weather station from four months (September through December 2016) between Mission IV and Mission V.
The dataset contains 32687 rows and 11 variables, which are about solar radiation, some weather information, and date information. Each data is observed by every five minutes.
The details of the columns are following.

Columns:

- UNIX time: Time date (seconds since Jan 1, 1970).
- Date: Date (yyyy-mm-dd format)
- Time: The local time of day (hh:mm:ss 24-hour format)
- Radiation: Solar radiation (watts per meter^2)
- Temperature: Temperature degrees (Fahrenheit)
- Pressure: Barometric pressure (Hg)
- Humidity: Humidity percent (%)
- WindDirection(Degrees): Wind direction degrees (degree)
- Speed: Wind speed (miles per hour)
- TimeSunRise: Sunrise Hawaii time (hh:mm:ss 24-hour format)
- TimeSunSet: Sunset Hawaii time (hh:mm:ss 24-hour format)

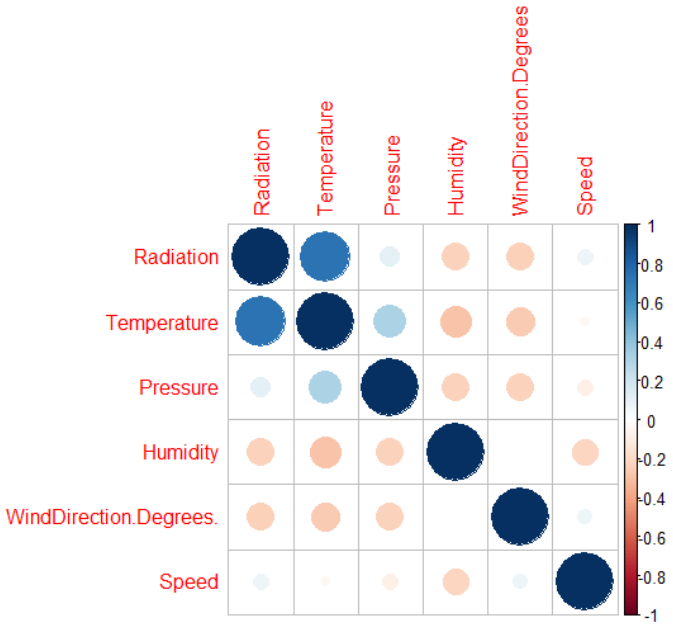
# 2 Exploratory Data Analysis

## 2.1 Data Summary

```
> summary(df)
    UNIXTime              Data               Time             Radiation        Temperature      Pressure
 Min.   :1472724008   Length:32686       Length:32686       Min.   :   1.11   Min.   :34.0   Min.   :30.19
 1st Qu.:1475546498   Class :character   Class :character   1st Qu.:   1.23   1st Qu.:46.0   1st Qu.:30.40
 Median :1478026070   Mode  :character   Mode  :character   Median :   2.66   Median :50.0   Median :30.43
 Mean   :1478047265                                         Mean   : 207.12   Mean   :51.1   Mean   :30.42
 3rd Qu.:1480480128                                         3rd Qu.: 354.24   3rd Qu.:55.0   3rd Qu.:30.46
 Max.   :1483264501                                         Max.   :1601.26   Max.   :71.0   Max.   :30.56
    Humidity      WindDirection.Degrees.     Speed          TimeSunRise         TimeSunSet
 Min.   :  8.00   Min.   :  0.09         Min.   : 0.000   Length:32686       Length:32686
 1st Qu.: 56.00   1st Qu.: 82.23         1st Qu.: 3.370   Class :character   Class :character
 Median : 85.00   Median :147.70         Median : 5.620   Mode  :character   Mode  :character
 Mean   : 75.02   Mean   :143.49         Mean   : 6.244
 3rd Qu.: 97.00   3rd Qu.:179.31         3rd Qu.: 7.870
 Max.   :103.00   Max.   :359.95         Max.   :40.500
```

The dataset is originally clean because there are no missing values. Only one variable, the WindDirection.Degrees., has been renamed to WindDirection to simplify. The range of Radiation is from 1.19 to 1601.26. This means that the value does not become 0 even when the nighttime, which is no sunshine. One variable that we must pay attention to is WindDirection because the value expresses the direction from 0 degrees to 360 degrees. Even though this is a continuous variable, it has cyclic characteristics because 360 degrees equal to 0 degrees.

## 2.2 Correlation Matrix




The data set has six continuous variables. The above correlation matrix shows that Temperature has a strong correlation with Radiation. Additionally, there is no strong correlation between the five predictor variables. Therefore, there is less likely multicollinearity.

2.3 Creating categorical variables

The following new three categorical variables are made for the deeper analysis.

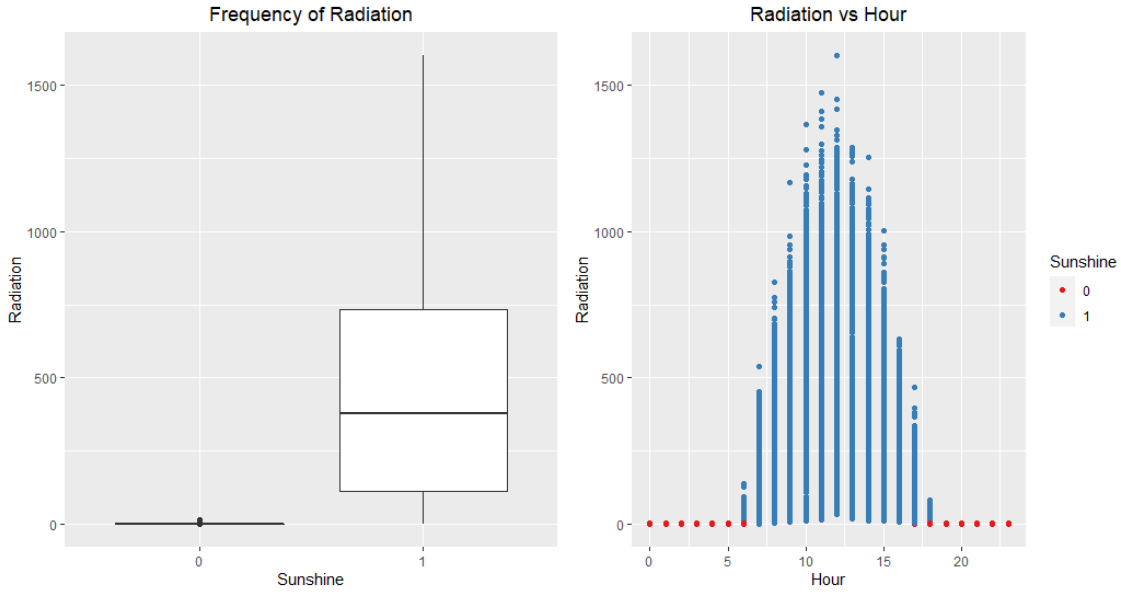
Hour: A hour indicator extracted from Time.

Wind_cat: A wind direction indicator that divides WindDirection into four types, such as 0 to 90, 90 to 180, 180 to 270, and 270 to 360.

Sunshine: A binominal variable that expresses 1 = daylight hour, 0 = non daylight hour. This is decided by whether during the time between sunset and sunrise or not.

2.4 Relationship between Radiation and Sunshine

In general, solar radiation is observed during daylight hours because the radiation is made by sunshine. Therefore, we will look at the cases during daylight hours and non-daylight hours. The following graphs are the relationship between Radiation and Sunshine variables.




The boxplot and scatter plot show that Radiation looks almost 0 when Sunshine is 0. Just in case, we will check the summary of both cases of the Sunshine.

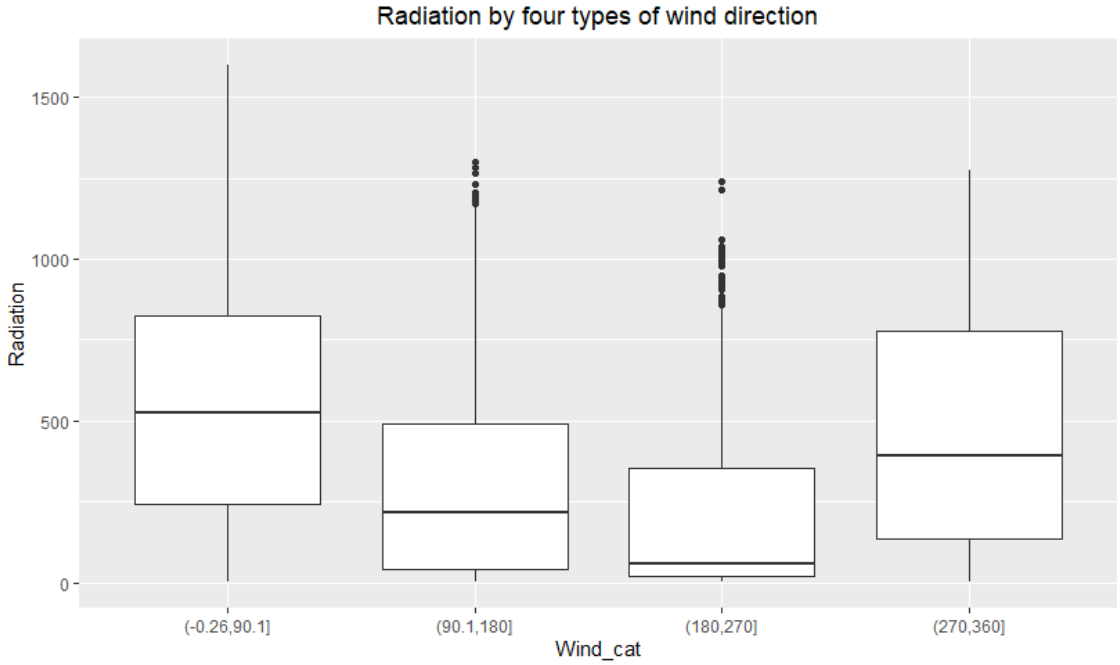
```
> summary(df[df$Sunshine==0,c(4,14)])
   Radiation      Sunshine
 Min.   : 1.11   0:17078
 1st Qu.: 1.21   1:    0
 Median : 1.23
 Mean   : 1.41
 3rd Qu.: 1.26
 Max.   :15.75
```

```
> summary(df[df$Sunshine==1,c(4,14)])
   Radiation       Sunshine
 Min.   :   1.19   0:    0
 1st Qu.: 113.10   1:15608
 Median : 379.43
 Mean   : 432.21
 3rd Qu.: 735.59
 Max.   :1601.26
```

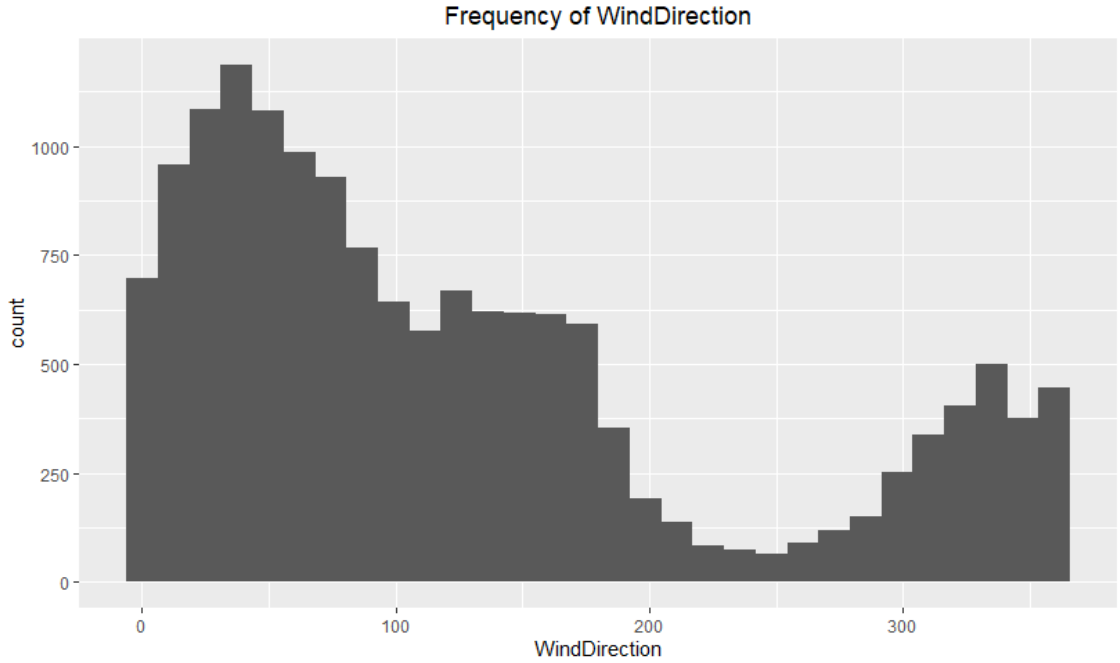
These summaries show that solar radiation is observed even during the non-daylight hour. However, the amount of Radiation is extremely lower than during daylight hour. The amount of Radiation is not enough to generate energy as solar generation. Therefore, we will focus on only the daylight hour data in this project.

2.5 Influence of Wind Direction

As mentioned earlier, it is difficult to deal with wind direction as numerical variables because 0 degrees equals 360 degrees. In order to understand the characteristics of Wind Direction, we will see the relationship between Radiation and Wind Direction using Wind_cat. Wind Direction originally does not indicate a specific direction, such as north or south. Therefore, Wind_cat is just made by separating four equal number ranges.

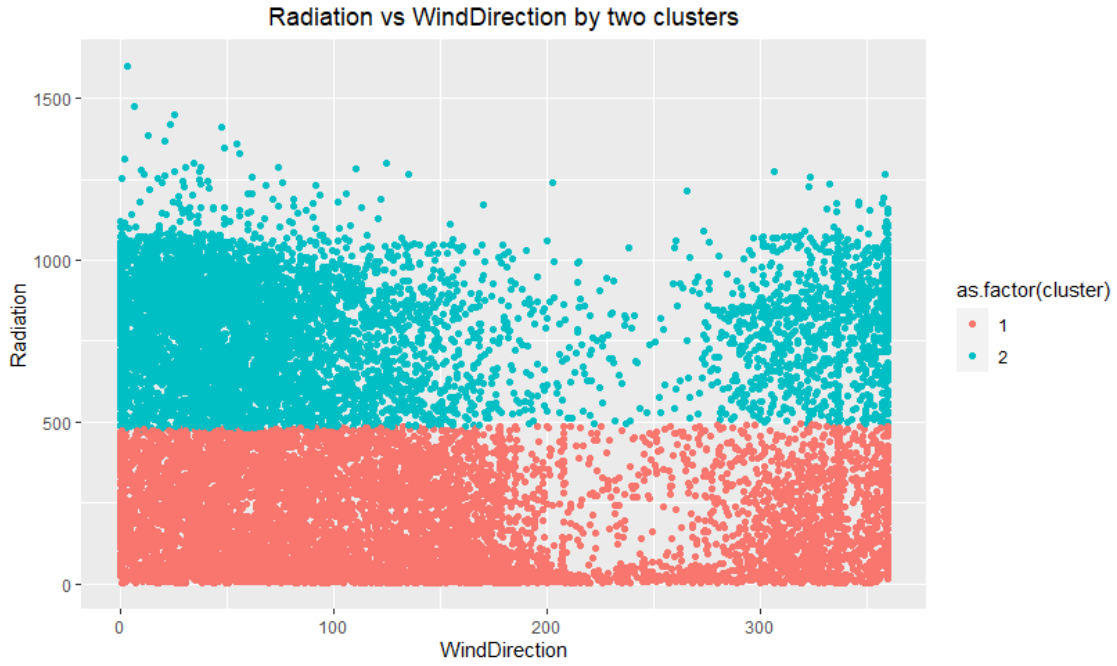


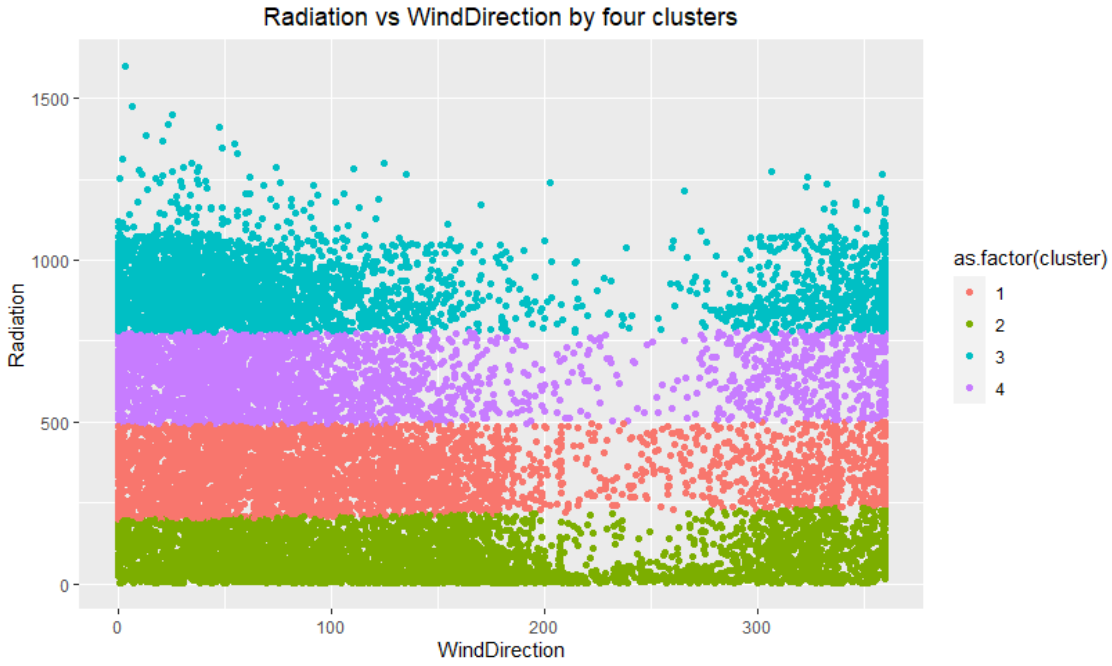

The boxplot seems that the range of Radiation depends on the Wind_cat. Especially, the first and fourth Wind_cat have a similar range, and the second and third also look similar. In addition to this, we will look at the frequency of Wind direction in the next.



This histogram is the frequency of the Wind Direction. Its frequencies largely differ in each Wind_cat. For example, there are more than a thousand counts in the first wind category, which has a wide range of Radiation distribution. In contrast, the third wind category does not have more than 250 counts. This difference could have affected the previous radiation distribution.

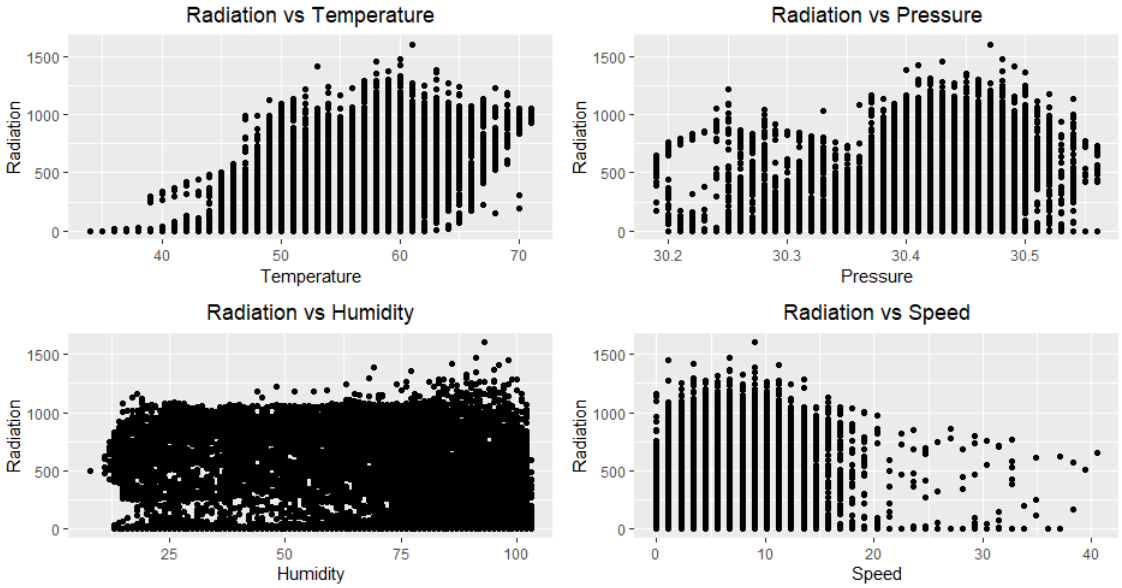
Then, we will look at clustering by using kmeans() function to see if there is a correlation between Radiation and Wind Direction. We will have two types of clustering, which have two clusters and four clusters.




In both cases, clusters do not depend on the Wind direction. Therefore, we will remove the Wind Direction because of the weakness of the correlation with radiation and the difficulty to deal with as a continuous variable.

2.6 Correlation of Radiation with four variables




As same as the result of the correlation matrix, Temperature seems to have a positive correlation with Radiation. In contrast, Speed looks like having a negative correlation with Radiation. On the other hand, Pressure and Humidity do not show a clear correlation in this graph.

# 3 Solar Radiation Prediction using Machine Learning Algorithms

## 3.1 Multivariable regression model 1

In this section, we will find the best model which can explain the value of Radiation well. The response variable, Radiation, is a continuous variable. Therefore, the linear regression model should be one of the best ways to figure it out.

### 3.1.1 Splitting the dataset

In order to test the accuracy of the linear regression model, the dataset has been split into two datasets. 60% percent of the dataset is a training dataset, and the other 40 percent of the data is a test dataset. This split was selected randomly.

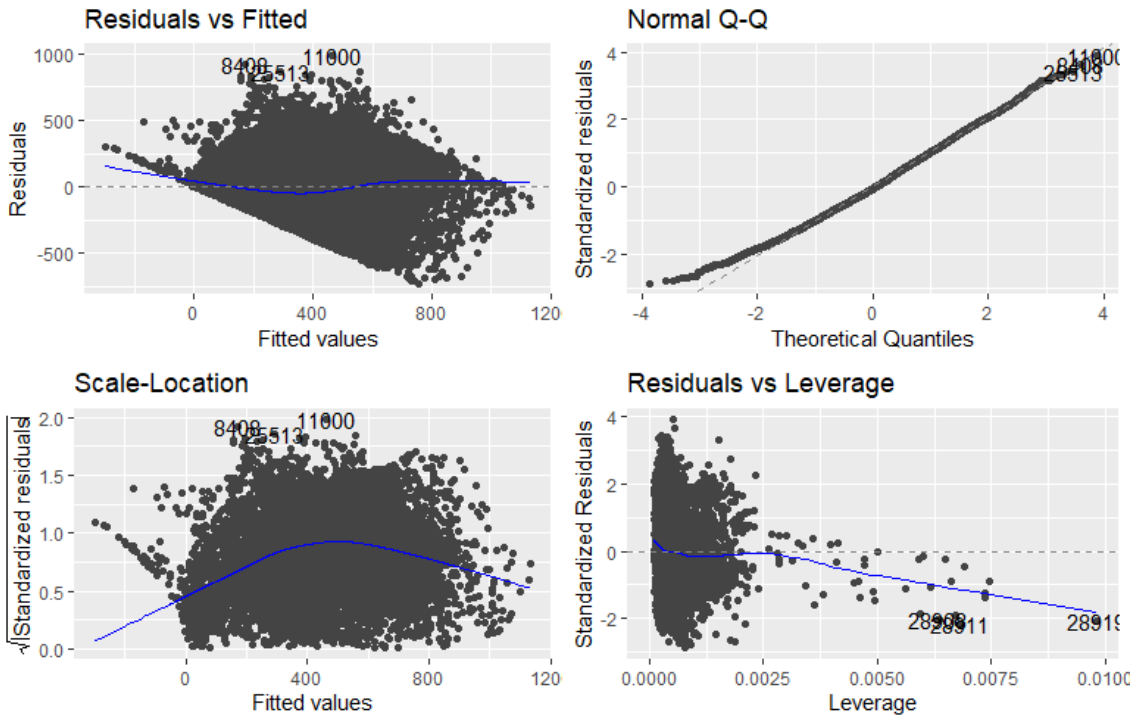
### 3.1.2 Creating a regression model 1 by training dataset.

The first model is calculated by Temperature, Pressure, Humidity, and Speed as predictors. The following is the result of the regression model.

```
Call:
lm(formula = Radiation ~ Temperature + Pressure + Humidity +
    Speed, data = train_df)

Residuals:
    Min      1Q  Median      3Q     Max
-728.98 -176.75  -13.77  173.60  990.19

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) -5521.3410  1626.5639  -3.394              0.00069 ***
Temperature    33.8812     0.5387  62.897 < 0.0000000000000002 ***
Pressure      133.3872    53.6296   2.487              0.01289 *
Humidity       -0.5493     0.1231  -4.463           0.00000819 ***
Speed          10.9950     0.7498  14.664 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 253.4 on 9360 degrees of freedom
Multiple R-squared:  0.4272,    Adjusted R-squared:  0.427
F-statistic:  1745 on 4 and 9360 DF,  p-value: < 0.00000000000000022
```

The P-values of all variables are less than 0.05, which are significant. Therefore, all four predictors are necessary for this model. Also, the Adjusted R-Squared is 0.427. This means that Radiation is 42.7% explainable by these four predictors. Unfortunately, this is not high enough to say good model.

3.1.3 Model1 Diagnostic plots



The above graphs are diagnostic plots of model 1. THE normal Q-Q plot looks good because residuals follow a straight line well. Additionally, Leverage doesn't have any influential observation. On the other hand, both Residual vs Fitted and Scale-Location shows a parabola line. This means that the weakness of the model 1.

### 3.1.4 Testing the model1 by test dataset

Finally, we will check the accuracy of the Model 1 by the Test dataset. The pearson r between predict Radiation and actual Radiation is 0.6646391. It is not a bad result because pearson r is over 0.05. However, it hard to say the model is good.

## 3.2 Multivariable regression model 2

Based on the result of the Model 1, we will find another model of the regression model. There is a possibility that the original data have unstable values because the observations are stationary measurements every 5 minutes. Therefore, we will create a second model created with a dataset that is observed on a daily average.

### 3.2.1 Transformation of the dataset

First of all, we aggregate the dataset by Date. The observation of the dataset for the second model is each date average or total. Taking account of each characteristic of variables, Radiation has been aggregated by total, and others have been aggregated by average. Moreover, there is a new variable named SunHours. SunHours is calculated by TimeSunSet -TimeSunRise. We will use it as a new predictor in the second model.

```
> summary(df_avg)
     Data           Total_Radiation    Avg_Temp      Avg_Pressure    Avg_Humidity      Avg_Speed         SunHours
 Length:15608       Min.   : 2145   Min.   :44.96   Min.   :30.23   Min.   : 19.56   Min.   : 2.695   Min.   :10.93
 Class :character   1st Qu.:43465   1st Qu.:51.77   1st Qu.:30.40   1st Qu.: 59.87   1st Qu.: 5.398   1st Qu.:11.05
 Mode  :character   Median :62183   Median :55.54   Median :30.43   Median : 80.50   Median : 6.076   Median :11.43
                    Mean   :58975   Mean   :55.07   Mean   :30.42   Mean   : 73.89   Mean   : 6.285   Mean   :11.52
                    3rd Qu.:75820   3rd Qu.:58.41   3rd Qu.:30.45   3rd Qu.: 93.08   3rd Qu.: 6.843   3rd Qu.:11.93
                    Max.   :92708   Max.   :64.45   Max.   :30.52   Max.   :101.97   Max.   :14.981   Max.   :12.52
```

### 3.2.2 Splitting the dataset

The same as 3.1.1, the new dataset has been split into two datasets. 60% percent of the dataset is a training dataset, and the other 40 percent of the data is a test dataset. This split was selected randomly.

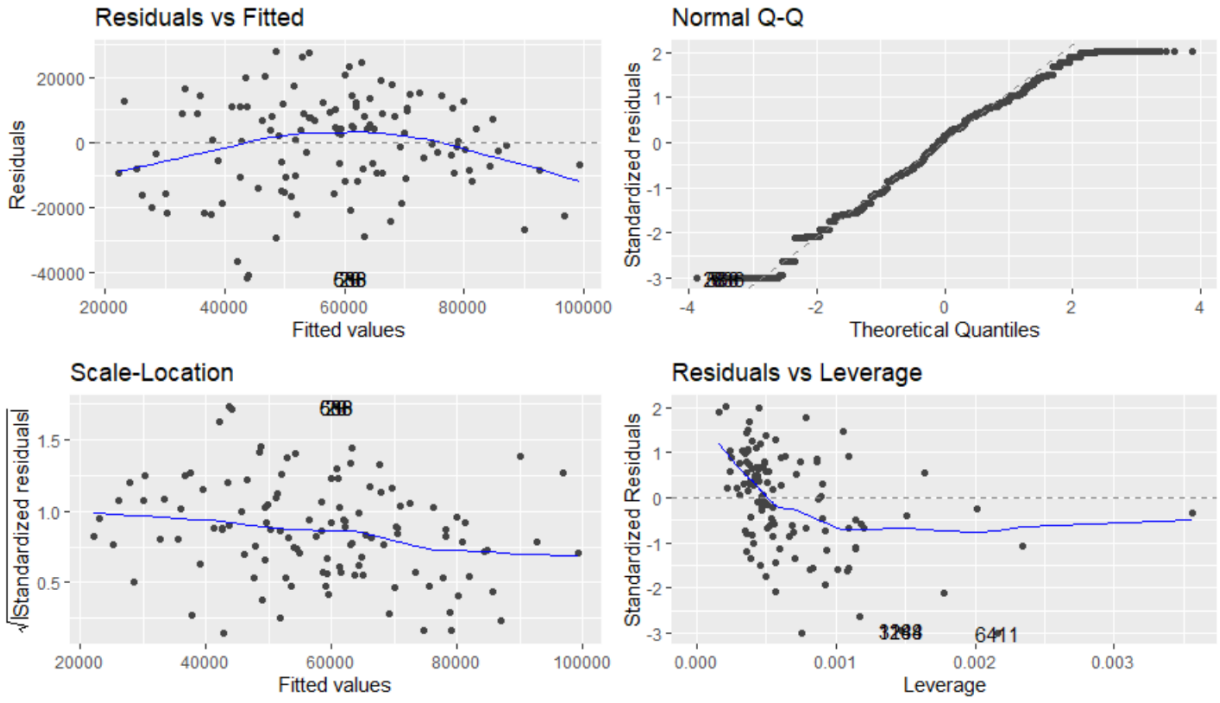
### 3.2.3 Creating a regression model 2 by training dataset

The response of the second model is Total_Radiation, and the predictors of the second model are Avg_Temp, Avg_Pressure, Avg_Humidity, Avg_Speed, and SunHours. The following is the result of the regression model.

```
Call:
lm(formula = Total_Radiation ~ Avg_Temp + Avg_Pressure + Avg_Humidity +
    Avg_Speed + SunHours, data = train_df_avg)

Residuals:
   Min     1Q Median     3Q    Max
-41547  -9539   1986  10441  27915

Coefficients:
               Estimate Std. Error t value             Pr(>|t|)
(Intercept)  -933873.01  107761.29  -8.666 < 0.0000000000000002 ***
Avg_Temp        2034.35      72.24  28.163 < 0.0000000000000002 ***
Avg_Pressure   26848.61    3566.62   7.528   0.0000000000000564 ***
Avg_Humidity    -274.49      11.92 -23.028 < 0.0000000000000002 ***
Avg_Speed       2998.30      92.14  32.542 < 0.0000000000000002 ***
SunHours        5661.17     461.58  12.265 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13870 on 9359 degrees of freedom
Multiple R-squared:  0.5936,    Adjusted R-squared:  0.5933
F-statistic:  2734 on 5 and 9359 DF,  p-value: < 0.00000000000000022
```

The P-values of all variables are less than 0.05, which are significant. Therefore, all five predictors are necessary for this model. Also, the Adjusted R-Squared is 0.5933. This means that Radiation is 59.3% explainable by these four predictors. Compared to the first model, the Adjusted R-Squared is 0.1663 higher. The reason why the second model is better might be adding the SunHour as a predictor, in addition to take the daily average in all predictors.

3.2.4 Model 2 Diagnostic plots



The previous graphs are diagnostic plots of model 2. In contrast to the result of model 1, the lines of both Residual vs Fitted and Scale-Location look more horizon. Additionally, the Normal Q-Q plot shows residuals follow a straight line well. Also, Leverage doesn't have any influential observation. Therefore, model 2 looks much better than model1.

### 3.2.5 Testing the model 2 by test dataset

In model 2. the pearson r between predict Total_Radiation and actual Total_Radiation is 0.7713145. This value is much higher than the model 1' result. Therefore, we can say model 2 is a better model than mode1.

## 4 Conclusion

In conclusion, Temperature, Pressure, Humidity, and Speed are valid predictors of Solar Radiation prediction. However, those predictors can only 42.7% of the Solar Radiation. To increase the accuracy of the prediction, it is better to predict total Solar Radiation per day by average Pressure, Humidity, Speed, and Sun hours. The total Solar Radiation can be almost 60 % explainable by these factors. However, this value is not enough to improve the stability of solar generation. As further research, we can find other effective predictors for Solar Radiation.