The impact of weather conditions on the PV system energy generation

Anju, Jose(N10434411)
IFN703/4 Assessment 3

Executive Summary

Photovoltaic (PV) systems have currently been extensively utilised by homeowners as an environmentally friendly supply of energy to satisfy the growing demand of electricity as it generates inexhaustible power by changing over the daylight into usable power as light. However, the energy generated from the PV system is subjected to totally different climatic conditions such as solar irradiance, temperature, pressure, humidity, windspeed, windspeed directions and so on. Hence, it is vital to study the effect of these environmental variables on the output generated from the PV system. This study proposes a applied mathematics called Multiple regression analysis to analyse the effect of different weather variables on the potency of the energy generated by the installed PV system in Australia/Brisbane over the year 2019 and the results says that solar radiance is highly correlated to the PV energy generation and the other variables in the analysis are less important in the PV output generation. This study helps in the analysis of the power generated from PV and thus benefit in to determine how much more generation capacity is needed during the extreme weather conditions, which finally helps the respective authorities to develop some alternative solution to generate the required amount of electricity during the extreme weather conditions.

Introduction

With technological and environmental advancement, individuals need more energy to establish a superior living environment. Burning conventional fossil fuels poses many environmental problems as well as health issues [1, 2]. The environmental problems can be land degradation, pollution, Emissions, Ocean acidification, so on and the health issues include asthma, bronchitis and other respiratory diseases. In order to deal with these complications renewable energies [3] such as solar energy, wind energy, tidal energy, biomass energy and so on are presented and these wellspring of energy are currently being utilized like never before. Of these, the photovoltaic business is quickly forming into one of the most encouraging business sectors in the field of sustainable power. Solar PV is the most currently used type of PV technology. However, other types of PV such as PV that can integrate into building, flexible PV are also being developed.

PV [4] is a device that creates power straightforwardly from daylight and is generally alluded to as "sunlight-based cells" [5], which are liberated from the utilization of ozone harming substances and petroleum products. Where solar cells are connected in series and in parallel to the current to increase the voltage and produce PV modules. The accessible intensity of one module is inadequate to flexibly more than one load. Therefore, numerous PV modules can be associated in arrangement or both to make a PV field to expand the yield current or voltage. Crystal line silicon PV modules [5] are broadly utilizes everywhere on the world. More up to date PV innovations are accessible today

with lower creation costs than conventional silicon-based modules, for example, formless silicon, cadmium telluride, and copper indium selenide.

Climate and other environmental factors such as temperature [6], dinghy [7], wind speed, geography, solar radiation affect PV module voltage and power output, so the efficiency of the solar system is not easily achieved [8-10] That is, climate change affects the temperature of the atmosphere and therefore it alters the energy produced by PV systems [11]. Therefore, it is important to differentiate between energy generated by the PV systems under different climatic conditions. Several methods have been proposed to analyse the effects of climate change on PV production. However, estimating the performance of PV produced under different climatic conditions has been little explored.

This paper examines the efficiency of the energy generated and the relation between the weather variables and the output of the PV system under different climatic conditions such as temperature, solar irradiance, pressure, humidity, windspeed and windspeed direction on production efficiency produced by PV systems using a statistical technique called Multiple regression analysis, in-order to analyse the environmental factors that is highly correlated to the PV generated energy and thus it is helpful in designing and selecting PV modules with appropriate sizes and specifications in relation with the climate of the area where the PV is being installed also this model helps in determining how much amount of energy is need in off seasons to meet the human needs. The result of the study shows that among the variables solar radiance has a high relation with the energy generation and all the other variables are least important in the solar energy generation and thus based on this analysis we can say that during the absence of sunlight the energy generation will get affected results in the generation of lesser energy from the PV system.

Regression analysis is a statistical method used to determine the relationship between independent variables (Predictor variables) and a dependent variable (Response variable) [12, 13]. The type of regression model always depends on the available dataset for the dependent variable and the type of model that produces the best fit. This study demonstrates a multi-regression model [14]on Australian PV data for the year 2019 associated with weather data from solcast API of Brisbane for the year 2019 to performs a coherent analysis to determine the relationship between PV systems and outputs generated by PV systems to determine the efficiency of output under different climatic conditions.

Literature Review

The advancement of solar energy system is treated as an important explanation to energy crises and environmental problems in today's world. However, long-term changes in climate, such as in other climatic conditions, also involve uncertainty in the response to changes in the solar system's radiation. Many exploration ventures dependent on different climatic boundaries have featured the impacts of environmental change on sun-oriented energy information and the execution of PV frameworks.

The approach proposed in [15] develops a regression model for solar power generation that uses 32-month weather forecasting data and considers clouds and temperature as variables. However, there are some drawbacks as it depends on the accuracy of the data provided, and some variables such as humidity are not included in the solar elevation analysis. A European study [16] used Eurocordex [17] with a model to generate photovoltaic energy to assess the effects of climate change on photovoltaic

energy, and the results show The change in the supply of solar photovoltaic energy should be in the range (-14 ٪) compared to estimates under current weather conditions, + 2 ٪), the largest reduction in the results indicating the northern countries. The temporary stability of power generation is not so much affected by future climate conditions, which also shows a slightly positive trend in the southern countries.

The method of [18] implements various methods such as binary trees for predicting solar radiation using descriptive statistics, factor analysis and clarity index and weather forecast data with extra-terrestrial solar radiation. Based on these data, they forecast global solar energy use three hours, one day ahead, and two days ahead. [19] are studying the impact of climate change on the operation of the photovoltaic system in Australia using historical solar energy data and Australian climate change scenarios, resulting in the prediction of future solar data by using morphing techniques. [20] The study proposed a framework for using solar data for weather forecasting, weather forecasting, and a combination of two data from South Korea's Yong Am Power Station, forecasting 36 hours in advance. However, in this study, the climate data is not considered only the previous solar energy data.

[21] With weather data such as temperature, humidity, wind speed, and wind direction, the ELM method (Extreme Learning Machine) was used to train the model and analyse hourly solar radiation on a horizontal surface. A study conducted by [22] has developed a precipitation method algorithm using data from the last years using clustering techniques to predict PV energy emissions. Using the Monte Carlo principle, a step-by-step MATLB-based algorithm is introduced [3] to detect the allowable power limit, interpreting weather conditions such as temperature and solar radiation as a function of cloud presence. This algorithm is used to evaluate the effect of climate change on the energy generated by the PV system.

The study [23] proposes a neural network model technique and linear regression model to estimate the output power of a PV system under soiling conditions. Both the models are trained and evaluated using actual monitoring data and the results says that it is possible to predict the maximum energy generation of soiled PV modules using these models. [24] demonstrates inter-year variability in wind and PV generation data by comparing different methods. However, this model only considers the influence of weather on PV output.

Another study [25] proposes algorithms to predict the production of photovoltaic systems based on weather classification such as clear sky, cloudy day, foggy day and rainy day, as well as reference vector machines (SVMs) . They developed a model to predict PV output for a single station a day in advance based on SVM, weather data, and historical output data. Another study [26] uses SVM and K-nearest neighbours (KNN) machine learning techniques to analyse how PV inference system data and daily weather classification data affect classification performance. The results show that SVM performs better with smaller sample scales, while KNN is sensitive to the length of the training dataset and provides higher accuracy than SVM. Various examinations and studies have been led to dissect the impacts of meteorological conditions on the exhibition of PV frameworks utilizing an assortment of measurable and AI strategies. The greatest test for some, specialists is information exactness and accessibility. This article proposes a strategy to research the impact of climate conditions on the effectiveness of the created intensity of a PV framework utilizing sun based investigation information and sun powered radiation utilizing a statistical technique called regression analysis, which says about the relation between the response and predictor variables.

Approach & Methodology

This study adopted a Multi-Regression analysis to explore the effect of different environmental parameters on the performance of the PV systems using solar analytics and weather data.

The Data

Three different datasets are used in this study.

- Solar Analytics Data

  Solar Analytics Historical data is used in this study. the solar analytics dataset is for 2019 period, at 5min interval, from 2019-01-01 00:00:00+00:00 to 2019-12-31 23:55:00+00:00 in Coordinated Universal Time (UTC) time format. The dataset contains information about 1000 customers in Australia.

States	No. of cutomers
ACT	2
NSW	251
QLD	229
SA	410
VIC	89
WA	14
TAS	5

Table 1. Distribution of customers in the dataset

  The information available for 1000 customers in the dataset is:

  - t_stamp_utc: solar analytics dataset is for 2019, at 5min intervals, from 2019-01-01 00:00:00+00:00 to 2019-12-31 23:55:00+00:00 in UTC format.
  - site_id: which is the customer code, where the PV system is installed for 1000 customers.
  - postcode: postcode of the customer location.
  - energy_(Wh): The energy generated from the installed PV systems during evry 5min for the whole year 2019.
  - voltage_max(V): the maximum voltage measured at the customer home during that 5 min period.
  - Voltage_min(V): the minimum voltage measured at the customer home during that 5 min period.

- Weather Data

  Historical record of weather data for the location Brisbane is downloaded from the solcast API toolkit for the year 2019. It contains both solar radiation and meteorological variables. Solcast offers a time granularity of 60min in UTC time format. The information about different weather parameters is available under this dataset and they are:

  - Period start & period end: weather dataset is for 2019, at 60min intervals, from 2019-01-01 00:00:00+00:00 to 2019-07-31 23:55:00+00:00 in UTC format.
  - AirTemp: Temperature (TEMP, °C)

- Dhi: Diffuse Horizontal Irradiance (DHI, W/m^2)
- Dni: Direct Normal Irradiance (DNI, W/m2)
- Ghi: Global Horizontal Irradiance (GHI, W/m^2)
- RelativeHumidity: Relative Humidity (RH, %)
- SurfacePressure: Surface Pressure (AP, hPa)
- WindDirection10m: Wind Direction (WD, °)
- WindSpeed10m: Wind Speed (WS, m/s)

- Solar Analytics Site Details

  Site Details dataset contains the information about the location where the solar panels are located. It includes:

  - Site_id: The location of the site where the solar panels is located
  - Postcode: Postcode of the area where that site is
  - State: State of that site_id
  - Timezone_id: Tells about the time zone that site comes under

Approach & Methodology

- **Data Pre-processing**
  - Data Loading & Filtering
    - Solar Data & Site Details: The dataset contains solar analytics information for the year 2019, at every 5 min intervals is loaded. From the 2019 data, January – July data is selected for the analysis as the number of observations for the whole is year beyond the capacity of the system. Then merge the solar data and site details to get more detailed information about the location of the solar panels in the solar analytics dataset. From the merged dataframe, based on time zone Australia/Brisbane data is selected, which gives information about 229 customers who are using solar panels and the unwanted columns such as state, site_id, timezone_id is removed from the dataframe for the analysis
    - Weather Data: Weather data for the period 2019 January to July, observed at every 60 min is loaded and removed the columns period end and period from the dataframe for the ease of analysis. Diffuse Horizontal Irradiance (Dhi)& Direct Normal Irradiance (Dni) comes under Global Horizontal Irradiance. So, we are not considering Dhi and Dni in our analysis.
  - Data Manipulation
    - Solar Data: The time in the dataset is converted into local Brisbane time as the time is given in UTC format and change the column name of time into date_time The observations in the Brisbane weather data is for every 60 min time intervals. In order to make the solar data into the same format, first we group on the basis of postcode and then get 60 min intervals observation from the 5 min interval data and calculated the average of the energy generated , minimum and maximum of the voltage generated based on the grouped data and thus we get the hourly data for the Brisbane region.

- Weather Data: The dataset is available in UTC timeframe, so it is converted into local Brisbane time and rename the time column name to date_time in order to match with the other dataset.

- Dealing with missing values

  The solar dataset contains missing values in minimum and maximum voltage measurement. Since, the analysis is to find the relation between the dependent and independent variables, the missing values are removed from the dataset.

- Merging the dataset

  In order to perform the regression analysis to examine the relation between the variables, merge the solar analytics dataset and the weather data and take the energy generated from the PV as the response variable for the analysis and all the other variables are considered as the predictor variables.

- **Exploratory Data Analysis**

  An exploratory data analysis is performed to study about the trend and pattern in the dataset and to understand how the data is distributed among the other parameters using visual methods. Here, an exploratory analysis is performed to find the relation between the energy generated from the PV with the other variables in the dataset

  A graph is plotted in order to determine, how the energy distribution is varying for each month in the given time period. The below graph shows the same.

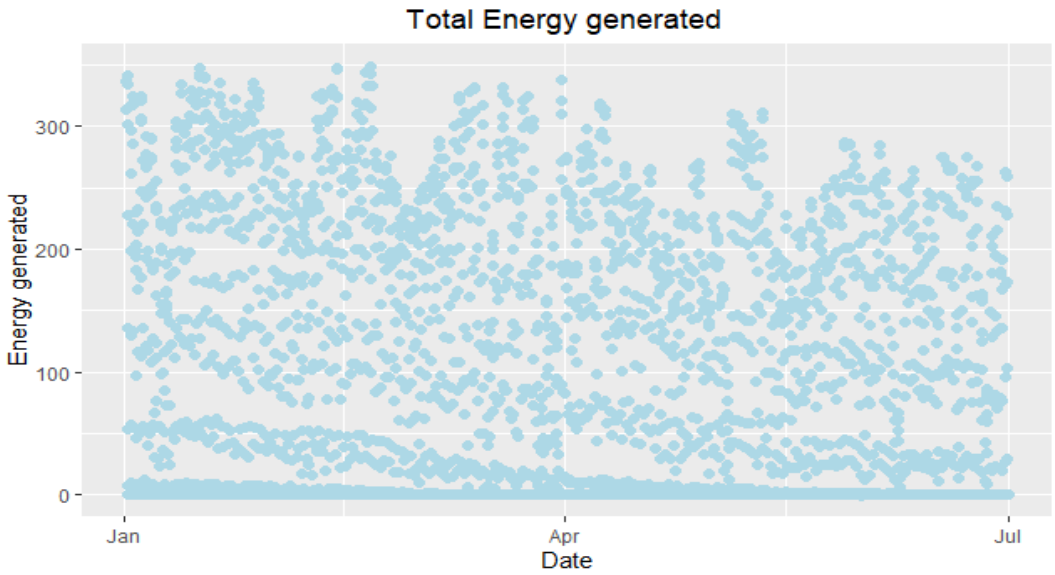


Figure 1. Total energy generated for each month

The above graph is too clumsy even though it says that more energy is generated during the month of January as it is showing observation for every hour. To visualise the plot in a detailed way, the hourly data is grouped into daily format and get the corresponding aggregate of the variables and visualise that. The below graph shows the energy generated per day during the given time period.

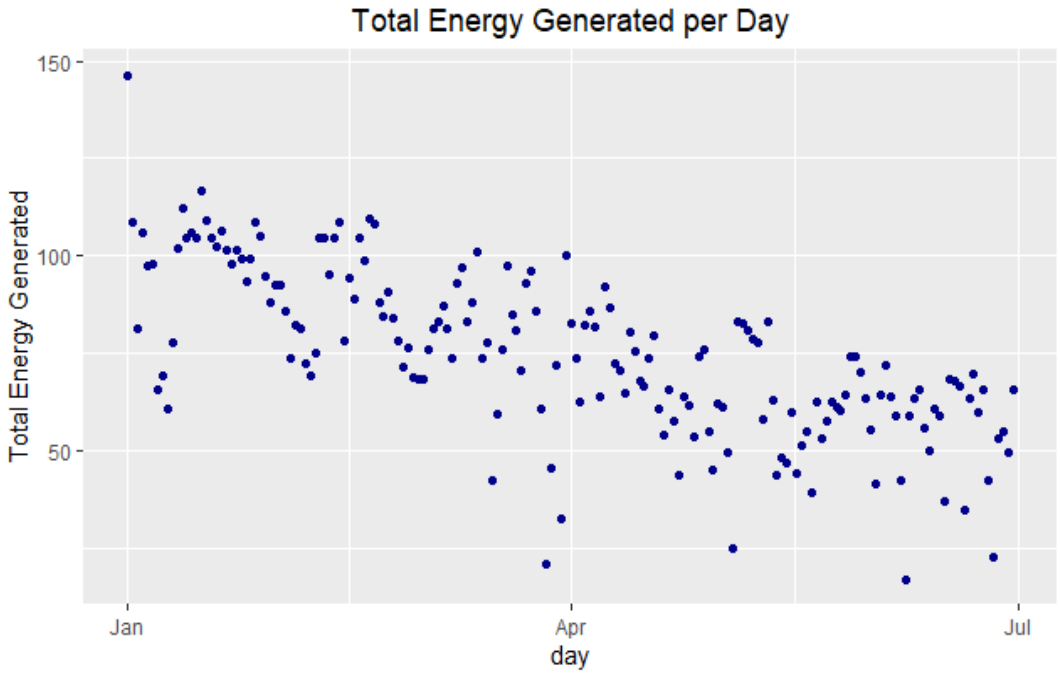


Figure 2. Total energy generated per day

The graph shows the total energy generated per day and it says that very high amount of energy is generated during the starting of January, which is the summer season and it get decreased eventually and during the winter, i.e., towards the end July the energy generated is less compared to the other days of the months. The energy distribution for the six months is shown in Figure.3.

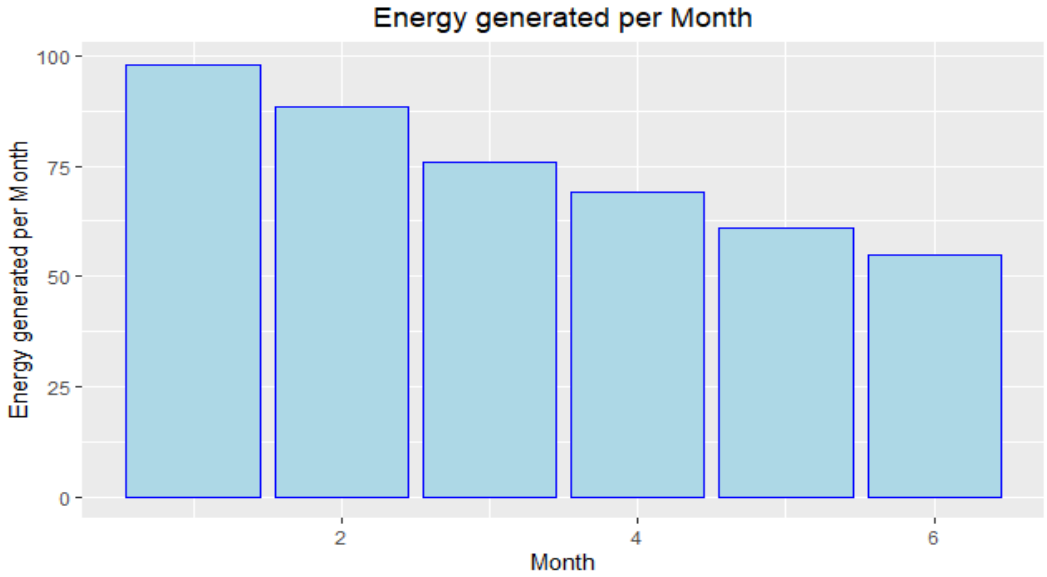


Figure 3. Energy generated per Month

Figure. 3 says that during the month of January more energy is generated, and July is the month where less energy is generated. Apart from that from the graph the energy generation is decreased by each month as the season get changed from summer to winter.

The below are the graphs which shows the relation between the energy generated and the other variables. It says, how each variable is related to the energy generation.

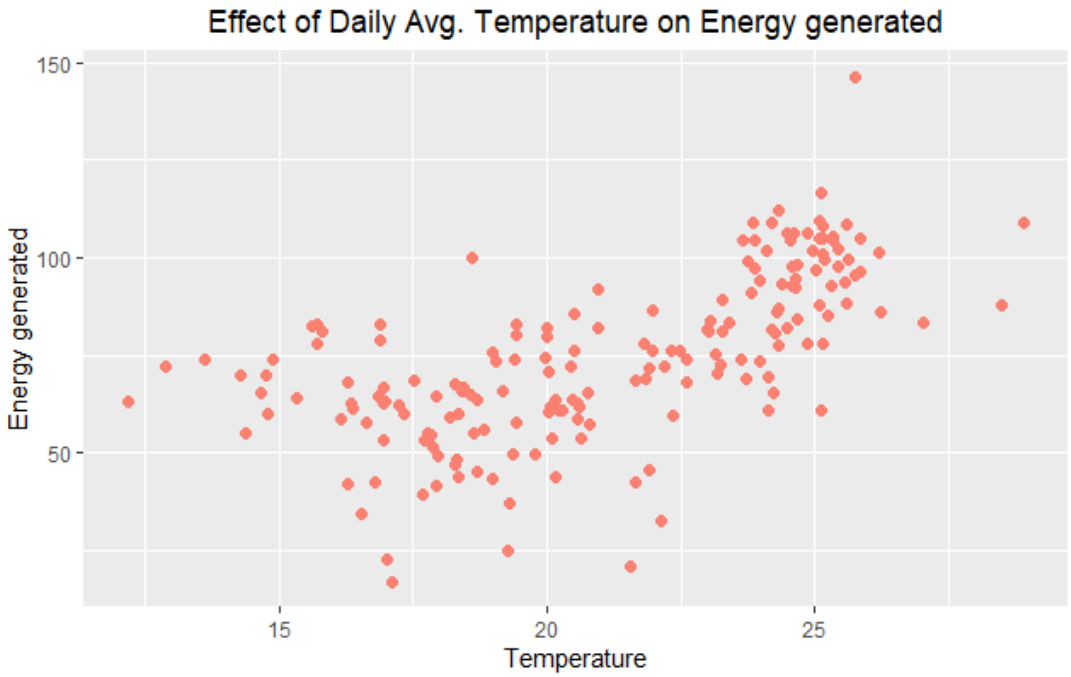


Figure 4. Energy generation Vs Temperature

The graph says that, when the temperature is less energy is generated and with an increase in temperature the energy generated also increases. So, we can say that, there is a linear relationship between the energy generation and temperature. Figure 5 shows the relation between the Global Horizontal irradiance and the daily average energy generation.

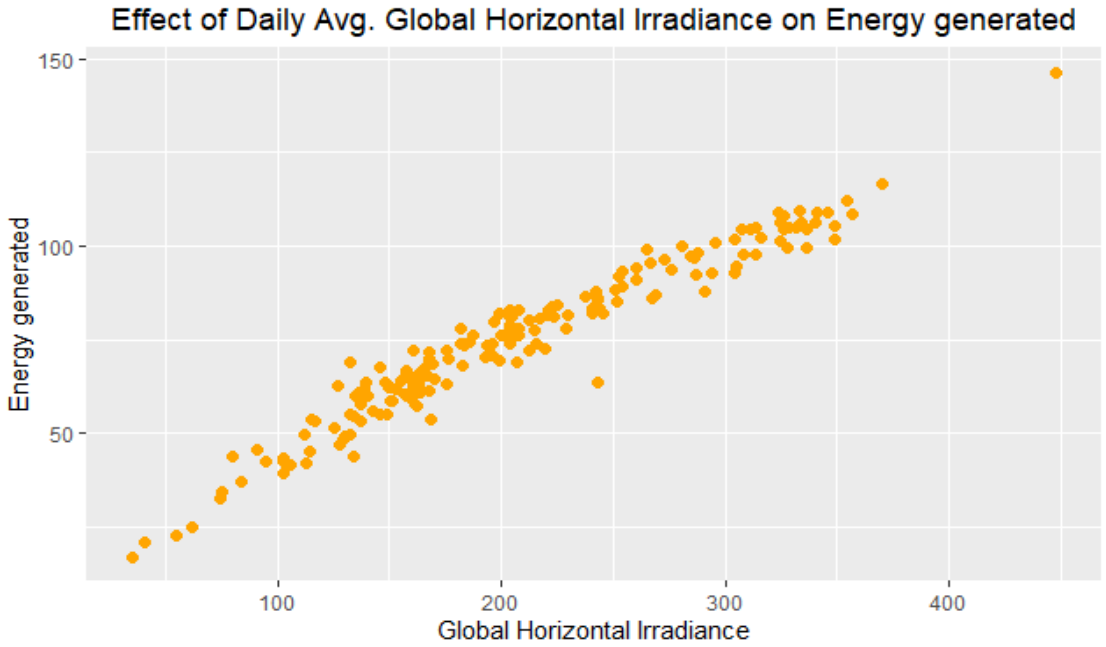


Figure 5. Global Horizontal Irradiance Vs Energy generation

The above figure states that when the GHI increase the energy generation also increases, which shows a linear relationship between the energy generation and solar radiation. From

this, the solar radiance has a great impact or a direct relationship in the energy generation from the PV systems. In order to analyse the relation between the energy generation and Humidity a graph is plotted below.

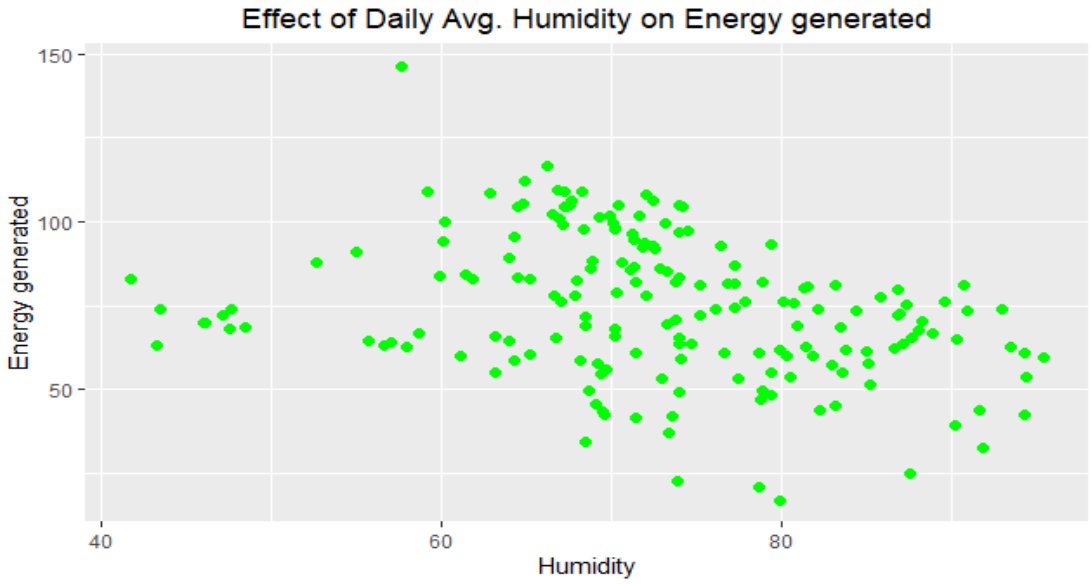


Figure 6. Humidity Vs Energy generated

From the above graph, there is no relation between the energy generated and humidity. The energy generation is between 50-100 for lesser and higher value of humidity. To determine the relation between the pressure and energy generated a graph is shown below.

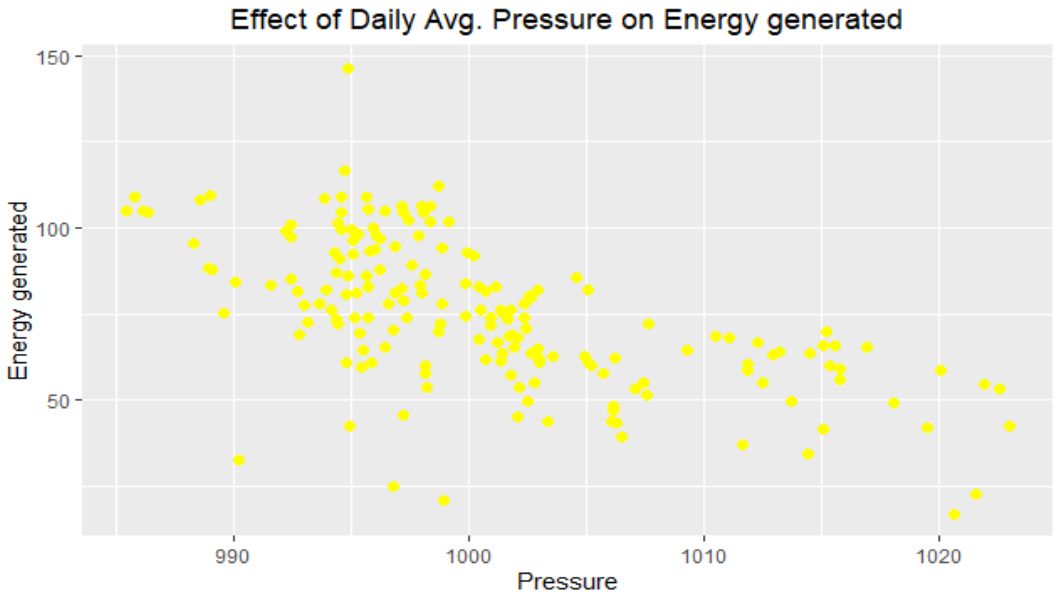


Figure 7. Energy generated Vs Pressure

The above figure says that, when the surface pressure is increasing the energy generated is decreasing. This shows an inverted relation between the pressure and energy generated from the PV system. The relation between the wind direction and energy generated is shown below

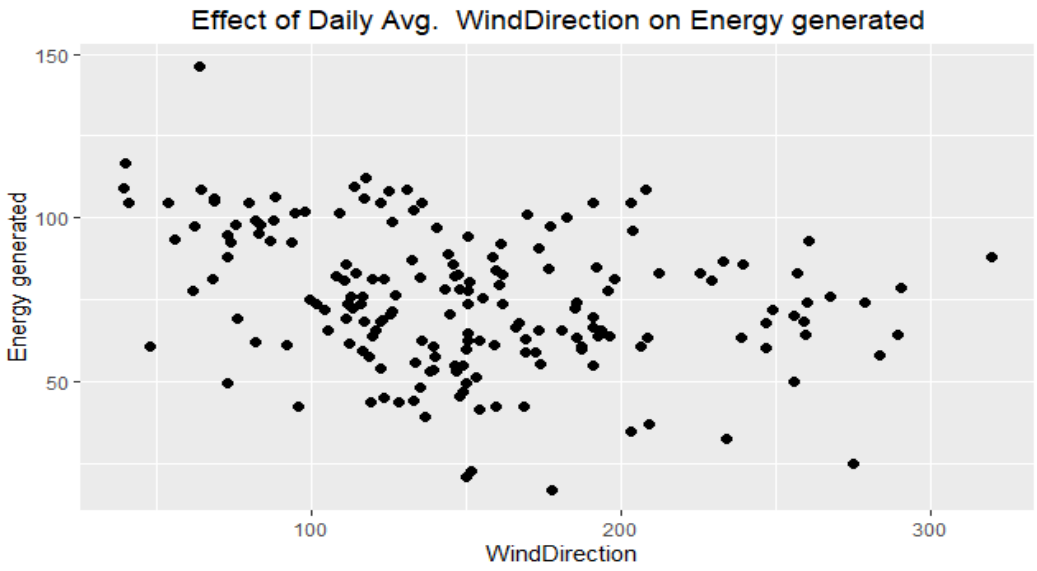


Figure 8. Energy generated Vs Wind Speed

The figure says that, winddirection has no relation with the energy generation. To analyse the relation between windspeed and energy generation a graph is displayed below.

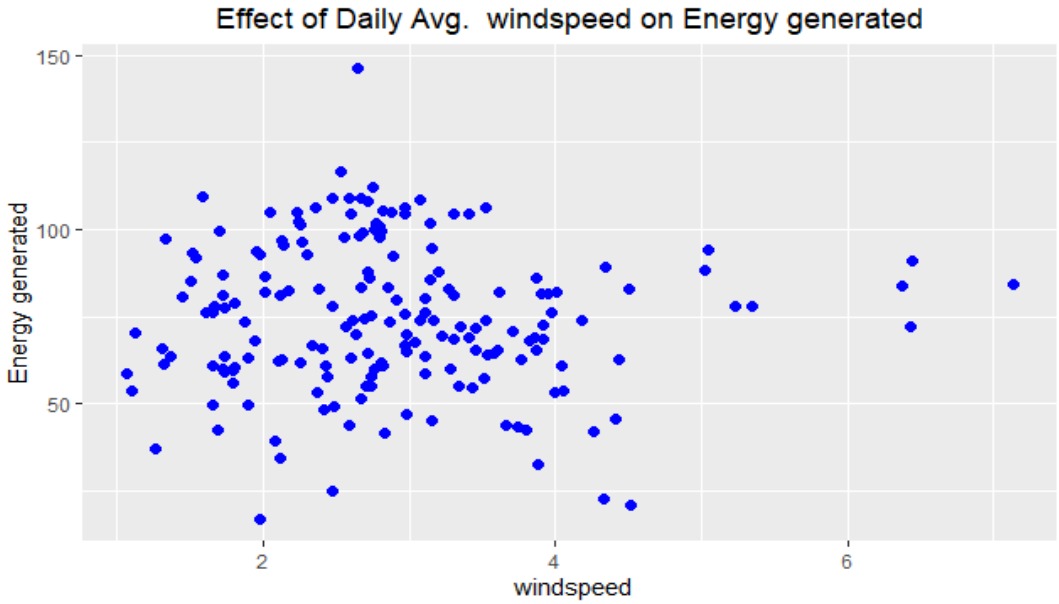


Figure 9. Energy Generated Vs Wind Speed

The figure explains that, when the wind speed is between the range 2 and 4 more energy is generated and it is almost same amount of energy is generated when the windspeed is 5 and above.

Based on all the graphs plotted, we can say that the Global Horizontal Irradiance is highly correlated with the energy generation from the PV system and all the other variables are least important in PV energy generation. That is, when the solar radiation is high, higher amount of energy will be generated from the solar panel and for the other climatic conditions less energy will be generated.

- **Model Building**

To analyse the effect of PV generation on weather variables, a Multi-Regression analysis is performed by considering the energy generated as the response variable and DHI, humidity, pressure, windspeed, windirection as the predictor variables. In order to determine the accuracy of the model on unseen dataset we split the data into test and train. 80% of the data is splitted as training set and the rest 20% is considered as the test data, which is regarded as an unseen dataset. The model is builded on the training data and evaluate using the testing data by finding the correlation accuracy between the model fit and the test data.

To determine the relation between each variable with the energy generation a simple linear regression is performed with the predictor variables. The mathematical expression for the linear regression model is defined below:

$$Y = b0 + b1*x + e$$

where, b0 is the y-intercept,

b1 is the coefficients associated with the predictor variables x1

e is the error term

```r
```{r}
fit1<- lm(`energy_(Wh)`~Ghi, data = train1)
fit2<- lm(`energy_(Wh)`~RelativeHumidity, data = train1)
fit3<- lm(`energy_(Wh)`~SurfacePressure, data = train1)
fit4<- lm(`energy_(Wh)`~WindDirection10m, data = train1)
fit5<- lm(`energy_(Wh)`~WindSpeed10m, data = train1)
fit6<- lm(`energy_(Wh)`~AirTemp, data = train1)
```
```

Figure 10. Linear model fits

The linear regression results give how each variable is related to the energy generation and it says that the Ghi is highly related with the energy generation and all the other variables are not having a direct relationship with the energy generation. Predictions are made against the fitted model and the test data also, the accuracy of the linear models using the test data is estimated, to see the difference between the fitted values against the data.

A Multi-Regression model is also built in order to determine the relation between the dependent and independent variables. Here a regression model is build using the training dataset and make predictions using the test dataset and finally compute the model accuracy. The mathematical expression for multiple regression model [13] is:

$$y = b0 + b1*x1 + b2*x2 + ... + bn*xn + e$$

where, b0 is the y-intercept,

b1, b2…. bn are the coefficients associated with the predictor variables x1, x2…. xn

e is the error term

```r
```{r}
###Fitting Models on the train data
fit7<- lm(`energy_(Wh)`~Ghi+AirTemp+RelativeHumidity+SurfacePressure+WindDirectio
n10m+WindSpeed10m, data = train1)
summary(fit7)
```
```

Figure 11. Multi -Regression model

The summary of the Multi regression model states that all the predictor variables are significant in the generation of the PV output. So, in order to determine the variables with higher significance a variable selection using stepwise analysis is performed.

```r
#Variable selection using stepwise aalysis
library(MASS)
stepAIC(fit7, trace = FALSE)$anova
```

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
`energy_(Wh)` ~ Ghi + AirTemp + RelativeHumidity + SurfacePressure +
    WindDirection10m + WindSpeed10m

Final Model:
`energy_(Wh)` ~ Ghi + AirTemp + RelativeHumidity + SurfacePressure +
    WindDirection10m + WindSpeed10m


  Step Df Deviance Resid. Df Resid. Dev      AIC
1                       3244    1623550 20213.79
```

Figure 12. Variable selection using stepwise analysis

The stepwise analysis result shows that all the predictor variables in the multi-regression model has significance. So, we are considering all the variables in building and fitting the model and the residuals for the fitted model is plotted to determine the non-linearity, error variances and outliers also, determined the R square value, which is the percentage of the dependent variable variation against the model.

Findings and Result

In simple linear regression model, we have only one predictor and one response variable, but in multiple-regression model we have a response variable and more than one predictor variable. We use lm() function to build the model. To analyse the relationship between the dependent and independent variables. The result of the simple Linear Regression analysis is given in the below table.

Linear Model	R Square	P-Value	Correlation Accuracy
`energy_(Wh)` ~ Ghi	0.9507	< 2.2e-16	0.9743054
`energy_(Wh)` ~ RelativeHumidity	0.3219	< 2.2e-16	0.5764313
`energy_(Wh)` ~ SurfacePressure	0.01686	1.068e-13	0.1781231
`energy_(Wh)`~WindDirection10m	0.003031	0.001688	0.2131693
`energy_(Wh)` ~ AirTemp	0.2947	< 2.2e-16	0.609013
`energy_(Wh)`~WindSpeed10m	0.02326	< 2.2e-16	0.2201767

Table 2. Linear Regression model Results

The result of the linear regression model defines that, The R Squared value is higher for the Global Horizontal Irradiance. For all the other predictors in the linear models have a smaller R Square value. Where the higher R Square value represents smaller difference between the observed data and the fitted values. Which means that higher the R Square, the better the model is. So, we can say that the GHI has a linear relation with the PV energy generation.

The residuals of the fitted model `energy_(Wh)` ~ Ghi is shown in Figure 12. Residual Vs Fitted is a scatter plot with residuals on the y axis and fitted values on the x axis. The Residual Vs Fitted plot is used to analyse the non-linearity or outliers in the data. This plot says that there are outliers in the

data and the values are scatted across the plot. So, we can conclude that residuals are not normally distributed in the data

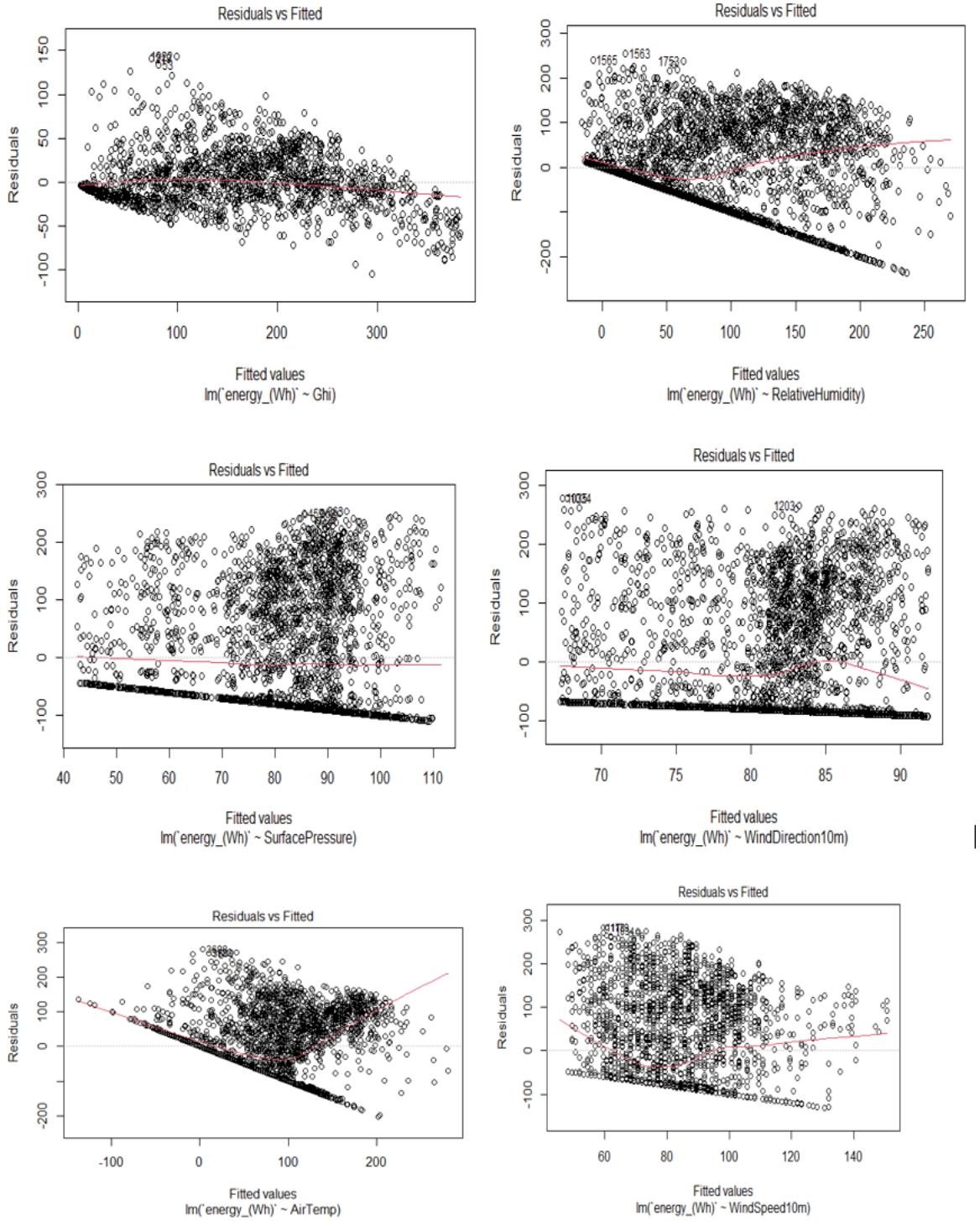


Figure 13. Residual Vs Fitted plot for the Linear Regression model

In determining the relation between the response variable `energy_(Wh)` and the other variables 'Ghi', 'RelativeHumidity', 'SurfacePressure',' WindDirection10m', 'AirTemp', 'WindSpeed10m' a

multi regression analysis is performed on the train data and evaluate on the test data. The result of the Multi Regression analysis is shown below:

Multi Regression Model	R square	P- Value	Correlation accuracy
`energy_(Wh)` ~ Ghi + AirTemp + RelativeHumidity + SurfacePressure + WindDirection10m + WindSpeed10m	0.9541	< 2.2e-16	0.9731785

Table 3. Multi Regression Model Result

The model resulted with an R square value of 0.9541, which says about the variance in the dependent variable `energy_(Wh)` which can be predicted from the independent variable Ghi, AirTemp, RelativeHumidity, SurfacePressure, WindDirection10m, WindSpeed10m. Hence, the value of R Square is higher, we can say that the model is good in predicting energy from the independent variables. Also, from the P-value, we can say that the predictor variables in the model is significant because the p-values are less that the common alpha level of 0.05. The correlation accuracy of the model results with 97.31%, says that the model is far good enough in prediction. The residual plot for the model is shown below

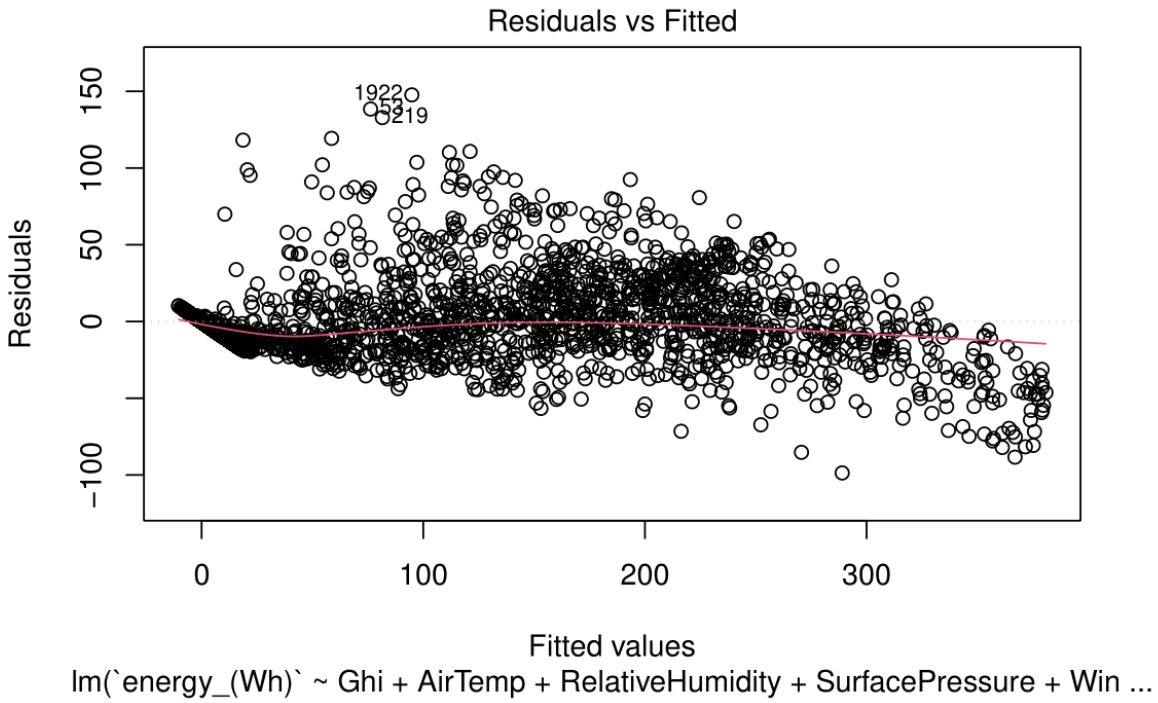


Figure 14. Residual Vs Fitted plot - Multi regression analysis

The graph shows that even though there are outliers in the model, most of the residual Vs Fitted values are laying around the centre line. So, from the R Square value, t residual plot and P-value of the fitted model, it is understood that the model is good in predicting from the independent values.

The simple linear regression as well as the multi regression analysis says that, solar radiation has great impact in the PV output generation. That is, when the solar radiance is high higher energy will be generated and vice versa for the other variables such as pressure, humidity, temperature.

Reflection

Weather condition is therefore subjected to the PV system output. A Multiple regression analysis is performed to understand the relation between the PV generated energy and with other environmental factors. Also, a simple linear regression is performed to analyse the relation between each variable with respect to the energy generated. The result of the simple regression says that, by considering weather variables individually Global Horizontal Irradiance has greater relation with R Square value 0.9507in the PV energy production and all the other variables are least important. The result for the Multiple regression analysis with predictors variables Ghi, pressure, temperature, humidity generates an R square value of 0.9541. From the results, we can conclude that the solar radiation has a greater impact in the energy generation and all the other variables has least importance in the PV energy generation. From the result, we can also predict the energy generation under different weather conditions, which will help to identify how much energy more is needed under worst climatic conditions such as Rainy season, snow period to meet the demand of electricity. These analysis will help the Engineers and technicians in the field of solar industry, to make decisions and develop strategies to develop solar panels with specifications and requirements that will work and generate electricity under all climatic conditions to meet the need of electricity in the future. The overall study helps me to understand more about what regression analysis is, how and where to use it. It gives me an idea about how to look for some data for the analysis, how to access data ethically from various site and how to use that data for different research purposes. Also, this study helps me to improve my data analysis skills like how to transform and analyse data by extracting answers to the research question.

The availability to access data was the greatest difficulty faced during the study. The solar data is for a period of 2019, at 5-minute interval for 1000 customers living at different locations and the details about the location such as postcode, region and time zone was available. To analyse the impact of the solar generation with weather variables, a weather dataset is also required. The Australian Bureau of Meteorology has got the postcode wise data. But downloading the data for 1000 customers postcode wise is not practically possible. Solcast API providing historical data, by creating an account.

Regression analysis for whole Brisbane shows that Solar radiance is the important weather variable in the energy generation. However, this can be different for some other countries like Finland, Sweden where the day light is very less throughout the year. A study based on different countries solar analytics data will give a more detailed explanation about how weather variable will affect the PV generation using Statistical and Machine Learning techniques. Numerous study was conducted to analyse the impact of weather variables in PV output was conducted and are undergoing. However, postcode wise analysis of the same is least explored. A future study based on this will explain about how weather variables effect the PV at different locations based on post code wise data.

References

[1] K. Kaygusuz, "Energy for sustainable development: key issues and challenges," *Energy Sources, Part B: Economics, Planning, and Policy,* vol. 2, no. 1, pp. 73-83, 2007.

[2] J. Peng, L. Lu, and H. Yang, "Review on life cycle assessment of energy payback and greenhouse gas emission of solar photovoltaic systems," *Renewable and sustainable energy reviews,* vol. 19, pp. 255-274, 2013.

[3] G. Boyle, *Renewable energy*. 2004.

[4] T. Ma, H. Yang, and L. Lu, "Solar photovoltaic system modeling and performance prediction," *Renewable and Sustainable Energy Reviews,* vol. 36, pp. 304-315, 2014.

[5] E. D. Dunlop, "Lifetime performance of crystalline silicon PV modules," in *3rd World Conference onPhotovoltaic Energy Conversion, 2003. Proceedings of*, 2003, vol. 3: IEEE, pp. 2927-2930.

[6] A. Hasan, J. Sarwar, and A. H. Shah, "Concentrated photovoltaic: A review of thermal aspects, challenges and opportunities," *Renewable and Sustainable Energy Reviews,* vol. 94, pp. 835-852, 2018.

[7] F. Belhachat and C. Larbes, "A review of global maximum power point tracking techniques of photovoltaic system under partial shading conditions," *Renewable and Sustainable Energy Reviews,* vol. 92, pp. 513-553, 2018.

[8] R. Zafarani, S. Eftekharnejad, and U. Patel, "Assessing the Utility of Weather Data for Photovoltaic Power Prediction," *arXiv preprint arXiv:1802.03913,* 2018.

[9] M. T. Chaichan and H. A. Kazem, "Experimental analysis of solar intensity on photovoltaic in hot and humid weather conditions," *International Journal of Scientific & Engineering Research,* vol. 7, no. 3, pp. 91-96, 2016.

[10] G. G. Kim *et al.*, "Prediction model for PV performance with correlation analysis of environmental variables," *IEEE Journal of Photovoltaics,* vol. 9, no. 3, pp. 832-841, 2019.

[11] J. A. Crook, L. A. Jones, P. M. Forster, and R. Crook, "Climate change impacts on future photovoltaic and concentrated solar power energy output," *Energy & Environmental Science,* vol. 4, no. 9, pp. 3101-3109, 2011.

[12] N. R. Draper and H. Smith, *Applied regression analysis*. John Wiley & Sons, 1998.

[13] S. Chatterjee and A. S. Hadi, *Regression analysis by example*. John Wiley & Sons, 2015.

[14] A. Sen and M. Srivastava, "Multiple regression," in *Regression Analysis*: Springer, 1990, pp. 28-59.

[15] J.-G. Kim, D.-H. Kim, W.-S. Yoo, J.-Y. Lee, and Y. B. Kim, "Daily prediction of solar power generation based on weather forecast information in Korea," *IET Renewable Power Generation,* vol. 11, no. 10, pp. 1268-1273, 2017.

[16] S. Jerez *et al.*, "The impact of climate change on photovoltaic power generation in Europe," *Nature communications,* vol. 6, no. 1, pp. 1-8, 2015.

[17] D. Jacob *et al.*, "EURO-CORDEX: new high-resolution climate change projections for European impact research," *Regional environmental change,* vol. 14, no. 2, pp. 563-578, 2014.

[18] F. Nomiyama, J. Asai, T. Murakami, and J. Murata, "A study on global solar radiation forecasting using weather forecast data," in *2011 IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2011: IEEE, pp. 1-4.

[19] G. Liu, X.-H. Tan, and M. Li, "Impacts of climate change on techno-economic performance of solar PV power systems: A case study in Australia," *Energy Procedia,* vol. 61, pp. 2588-2591, 2014.

[20] B. Carrera and K. Kim, "Comparison Analysis of Machine Learning Techniques for Photovoltaic Prediction Using Weather Sensor Data," *Sensors,* vol. 20, no. 11, p. 3129, 2020.

[21] I. Abadi and A. Soeprijanto, "Extreme learning machine approach to estimate hourly solar radiation on horizontal surface (PV) in Surabaya-East java," in *2014 The 1st International*

Conference on Information Technology, Computer, and Electrical Engineering, 2014: IEEE, pp. 372-376.

[22] M.-C. Kang, J.-M. Sohn, J.-y. Park, S.-K. Lee, and Y.-T. Yoon, "Development of algorithm for day ahead PV generation forecasting using data mining method," in *2011 IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2011: IEEE, pp. 1-4.

[23] S. Shapsough, R. Dhaouadi, and I. Zualkernan, "Using linear regression and back propagation neural networks to predict performance of soiled PV modules," *Procedia Computer Science,* vol. 155, pp. 463-470, 2019.

[24] S. Pfenninger, "Dealing with multiple decades of hourly wind and PV time series in energy models: A comparison of methods to reduce time resolution and the planning implications of inter-annual variability," *Applied energy,* vol. 197, pp. 1-13, 2017.

[25] J. Shi, W.-J. Lee, Y. Liu, Y. Yang, and P. Wang, "Forecasting power output of photovoltaic systems based on weather classification and support vector machines," *IEEE Transactions on Industry Applications,* vol. 48, no. 3, pp. 1064-1069, 2012.

[26] F. Wang, Z. Zhen, B. Wang, and Z. Mi, "Comparative study on KNN and SVM based weather classification models for day ahead short term solar PV power forecasting," *Applied Sciences,* vol. 8, no. 1, p. 28, 2018.

Appendix 1: Supplementary Information

Loading the weather data

```
require(data.table)
```

```
## Loading required package: data.table
```

```
setwd("D:/Sem 2, 2020/IFN704/Data")
#create a list of the files from your target directory
file_list <- list.files(path="D:/sem 2, 2020/IFN704/Data")

#initiate a blank data frame, each iteration of the loop will append the data from the given file to th
dataset <- data.frame()

#had to specify columns to get rid of the total column
for (i in 1:length(file_list)){
  temp_data <- fread(file_list[i], stringsAsFactors = F) #read in files using the fread function from t
  dataset <- rbindlist(list(dataset, temp_data), use.names = T) #for each iteration, bind the new data
}

#Loading the site details
site_details <-  read.csv("D:/sem 2, 2020/IFN704/site_details.csv")
```

Data Pre-Processing

```
#merging site details and solar data
solar_data <- merge(dataset,site_details,by="site_id")

#Extracting the Australia/Brisbane data
bribane_solar_data <- subset(solar_data, timezone_id == "Australia/Brisbane")

#missing values
brisbane_solar_data <- na.omit(bribane_solar_data)

#converting UTC time to local time
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:data.table':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
brisbane_solar_data$brisbane_localtime <- as_datetime(brisbane_solar_data$t_stamp_utc, tz = "Australia/
#Rearranging the columns
brisbane_solar_data <- brisbane_solar_data[, c(1,2,9,3,4,5,6,7,8)]
brisbane_solar_data$t_stamp_utc <- NULL
bris_post <- brisbane_solar_data
```

Sine the weather data availableis based on the local brisbane we are taking the aggregate of the solar data to make the dataset into Brisbane data instead of taking every postcodes.

```
library(plyr); library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data_brisbane <-bris_post %>%
  group_by((DeviceTime = cut(brisbane_localtime, breaks="hour")), postcode) %>%
  summarize(`energy_(Wh)` = mean(`energy_(Wh)`),
            `voltage_max_(V)`=max(`voltage_max_(V)`),
            `voltage_min_(V)`=min(`voltage_min_(V)`))
```

```
## `summarise()` regrouping output by '(DeviceTime = cut(brisbane_localtime, breaks = "hour"))' (overri
```

```
dataset_brisbane <- rename(data_brisbane, date_time = 1)
class(dataset_brisbane$date_time)
```

```
## [1] "factor"
```

```
data_hourly <- dataset_brisbane %>%
  mutate(DATE = as.Date(date_time, format = "%Y/%m/%d %H:%M:%S")) %>%
  summarize(`energy_(Wh)` = mean(`energy_(Wh)`),
            `voltage_max_(V)`=max(`voltage_max_(V)`),
            `voltage_min_(V)`=min(`voltage_min_(V)`))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
data_hourly <- head(data_hourly,-10)
data <- data_hourly
```

Loading weather data

```
library(lubridate)
solcast <-  read.csv("D:/sem 2, 2020/IFN704/solcast_data.csv")
solcast$PeriodStart <- as_datetime(solcast$PeriodStart, tz = "Australia/Brisbane")
```

```
## Date in ISO8601 format; converting timezone from UTC to "Australia/Brisbane".
```

```
colnames(solcast)
```

```
##  [1] "PeriodEnd"        "PeriodStart"      "Period"           "AirTemp"
##  [5] "Dhi"              "Dni"              "Ghi"              "RelativeHumidity"
##  [9] "SurfacePressure"  "WindDirection10m" "WindSpeed10m"
```

```
#Removing unused columns
solcast_data <- solcast[-c(1,3)]

#Renaming date_time column
names(solcast_data)[names(solcast_data) == "PeriodStart"] <- "date_time"
```

Merging the two dataframes

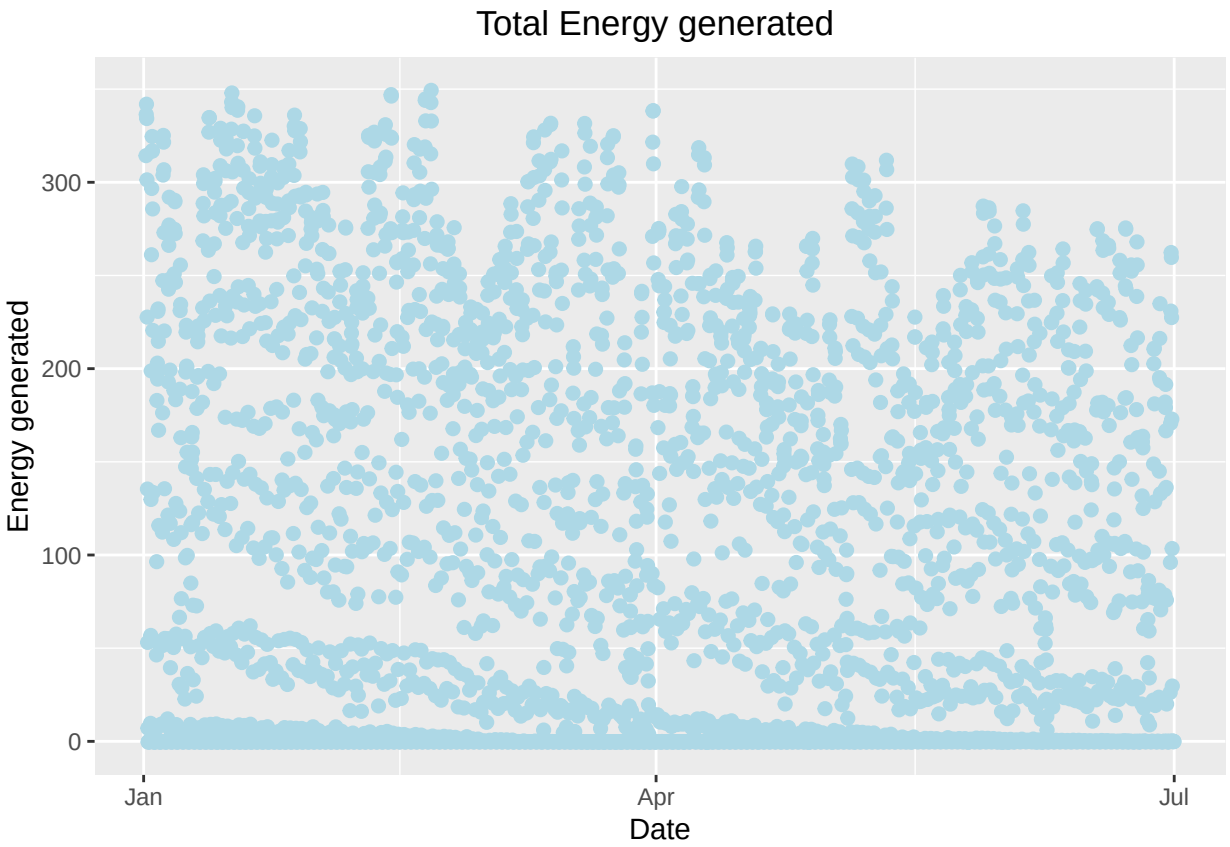
```
#merging weather data and solcast data
data_hourly$date_time <- as_datetime(data_hourly$date_time, tz = "Australia/Brisbane")
#merging weather data and solcast data
solar_weather <- merge(data_hourly,solcast_data,by="date_time")
```

Exploratory Data Analysis

Exploratory data analysis is performed to analyze the relation between the variables and the energy generated

```
library(ggplot2)
#solar_weather$date_time <- as_datetime(solar_weather$date_time, format= "%Y-%m-%d %H:%M:%S")
data1 <- solar_weather
data1$day <- as.Date(data1$date_time, tz="Australia/Brisbane",format="%Y-%m-%d")


ggplot(data1)+
  geom_point( aes(x = date_time, y = `energy_(Wh)`), colour = "lightblue", size = 2)+
  xlab("Date")+
  ylab("Energy generated")+
  ggtitle("Total Energy generated ")+
  theme(plot.title = element_text(hjust = 0.5))
```




```
# Calculating and plotting the Energy generated per day to see t he daily differences in the energy gen
library(dplyr)

colnames(data1)
```

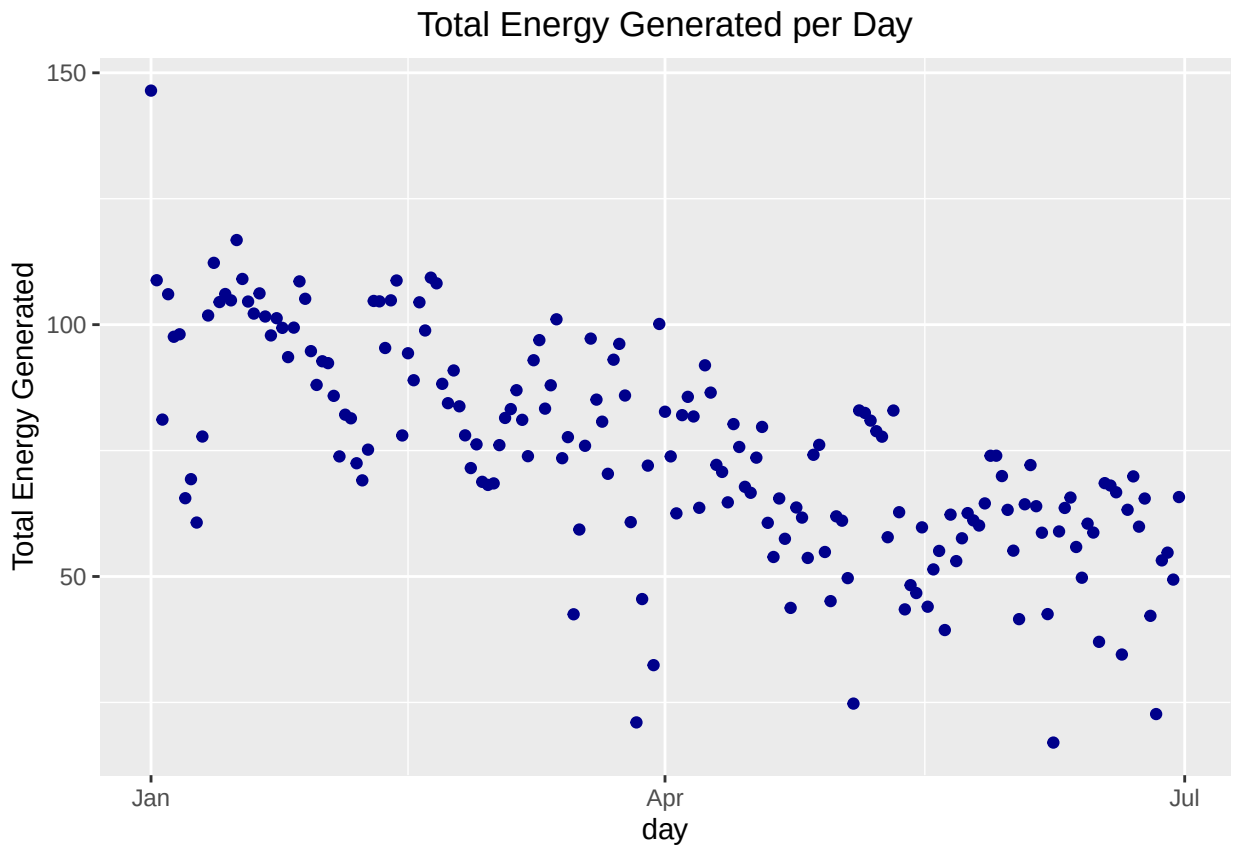
```
##  [1] "date_time"        "energy_(Wh)"      "voltage_max_(V)"  "voltage_min_(V)"
##  [5] "AirTemp"          "Dhi"              "Dni"              "Ghi"
##  [9] "RelativeHumidity" "SurfacePressure"  "WindDirection10m" "WindSpeed10m"
## [13] "day"
```

```
by_day <- data1 %>%
  group_by(day) %>%
  summarize(`energy_(Wh)` = mean(`energy_(Wh)`),
            `voltage_max_(V)`=max(`voltage_max_(V)`),
            `voltage_min_(V)`=min(`voltage_min_(V)`),
            Temp = mean(AirTemp),
            DHI = mean(Dhi),
            DNI = mean (Dni),
            GHI = mean(Ghi),
            Humidity=mean(RelativeHumidity),
            Pressure=mean(SurfacePressure),
            WindDirection=mean(WindDirection10m),
            windspeed=mean(WindSpeed10m))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
ggplot(by_day)+
  geom_point(mapping = aes(x = day, y = `energy_(Wh)`), colour = "darkblue")+
```

```
  ylab("Total Energy Generated")+
  ggtitle("Total Energy Generated per Day")+
  theme(plot.title = element_text(hjust = 0.5))
```



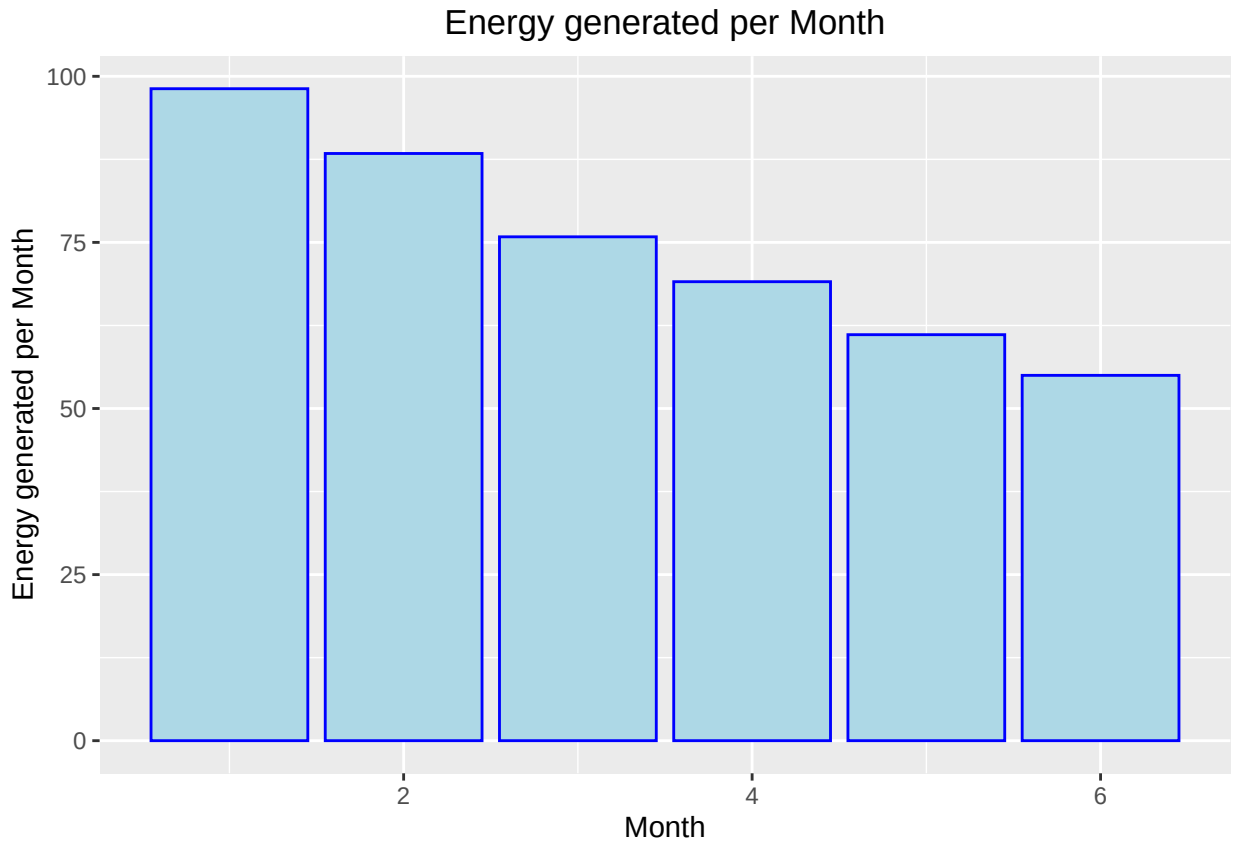

```
#Calculating and plotting Energy Generated per Month to perform a monthy analysis on the data
data1$Month <- month(data1$day)

month_rad <- data1 %>%
  group_by(Month) %>%
  summarize(`energy_(Wh)` = mean(`energy_(Wh)`),
            `voltage_max_(V)`=max(`voltage_max_(V)`),
            `voltage_min_(V)`=min(`voltage_min_(V)`),
            Temp = mean(AirTemp),
            DHI = mean(Dhi),
            DNI = mean (Dni),
            GHI = mean(Ghi),
            Humidity=mean(RelativeHumidity),
            Pressure=mean(SurfacePressure),
            WindDirection=mean(WindDirection10m),
            windspeed=mean(WindSpeed10m))
```

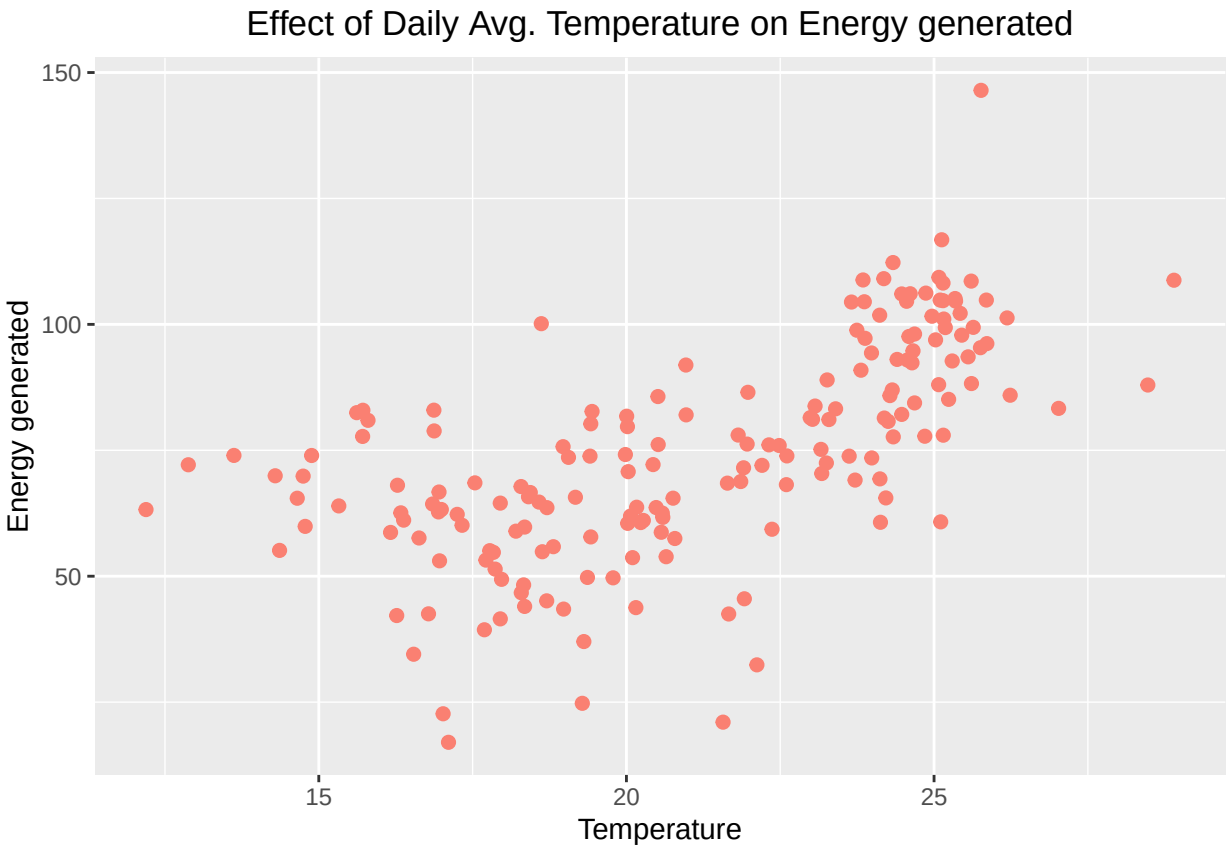
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
ggplot(month_rad, aes(x = Month, y = `energy_(Wh)`))+
  geom_bar(stat = "identity", colour = "blue", fill = "lightblue")+
  xlab("Month")+
  ylab("Energy generated per Month")+
```

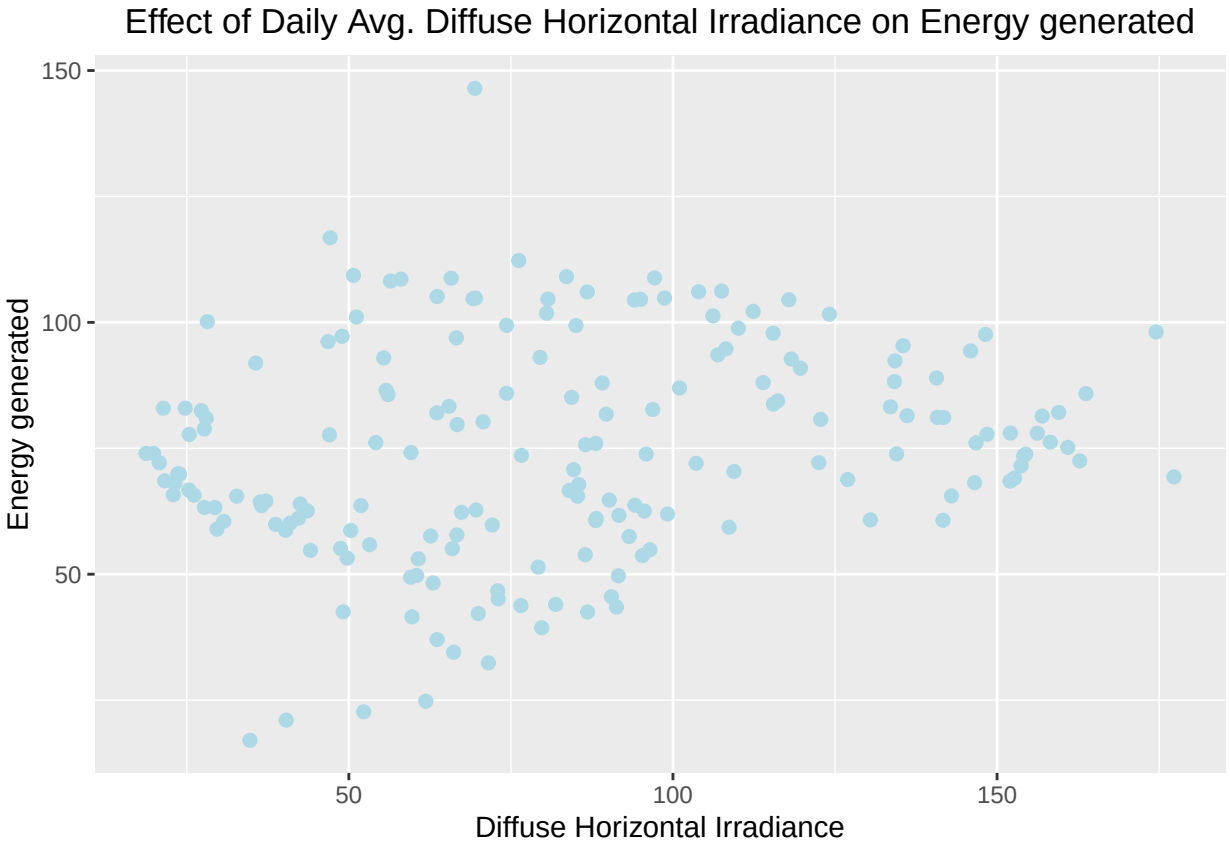
```
  ggtitle("Energy generated per Month")+
  theme(plot.title = element_text(hjust = 0.5))
```



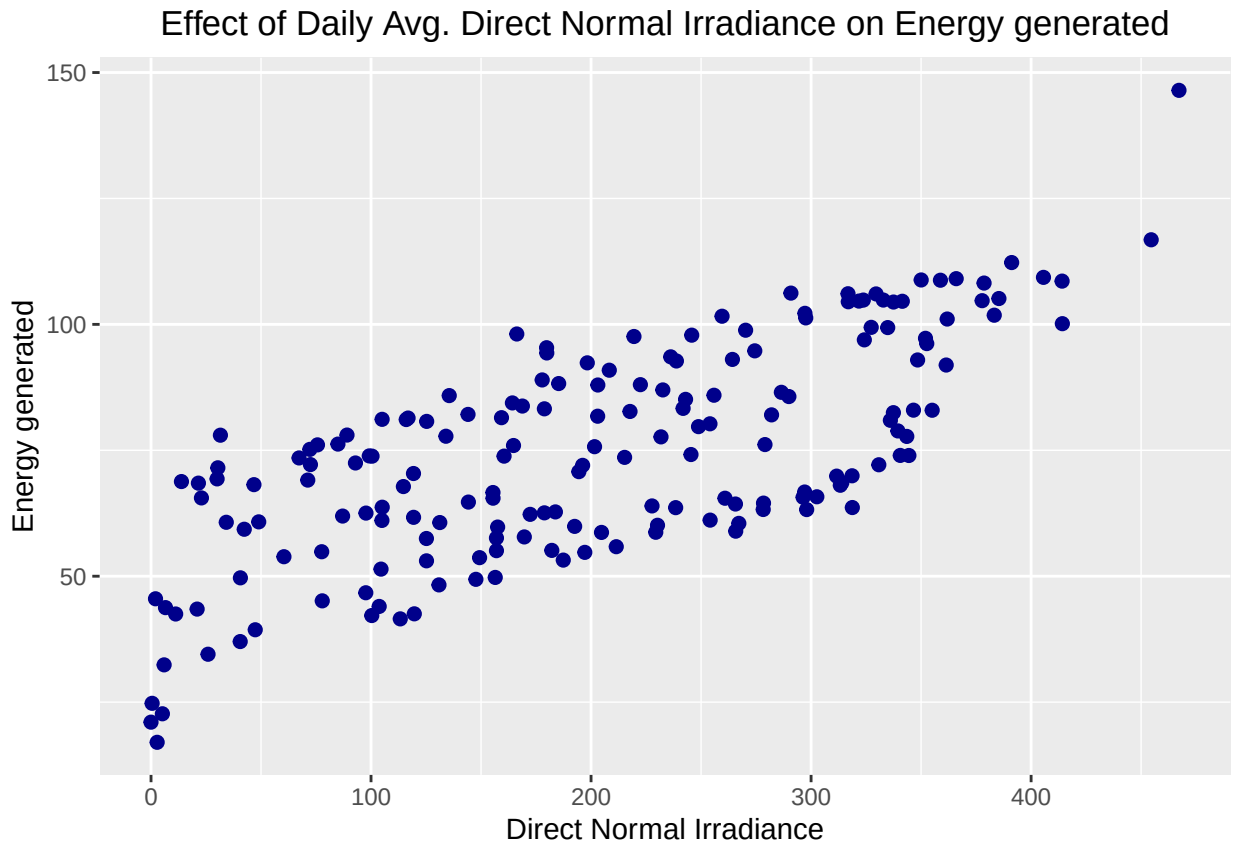

```
#Effect of weather variables with the energy generated
ggplot(by_day)+
  geom_point(mapping = aes(x = Temp, y = `energy_(Wh)`), colour = "salmon", size = 2)+
  ylab("Energy generated")+
  xlab("Temperature")+
  ggtitle("Effect of Daily Avg. Temperature on Energy generated")+
  theme(plot.title = element_text(hjust = 0.5))
```



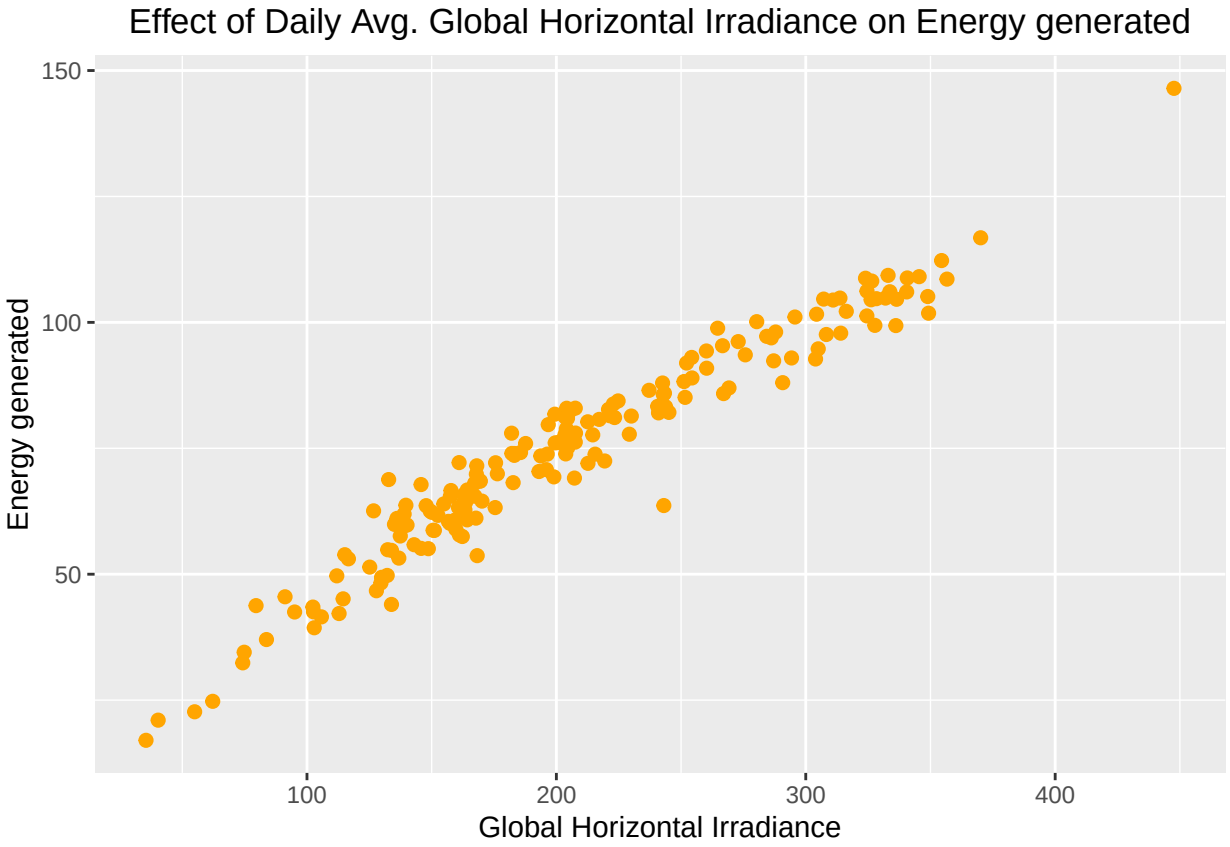

```
ggplot(by_day)+
  geom_point(mapping = aes(x = DHI, y = `energy_(Wh)`), colour = "lightblue", size = 2)+
  ylab("Energy generated")+
  xlab("Diffuse Horizontal Irradiance")+
  ggtitle("Effect of Daily Avg. Diffuse Horizontal Irradiance on Energy generated")+
  theme(plot.title = element_text(hjust = 0.5))
```



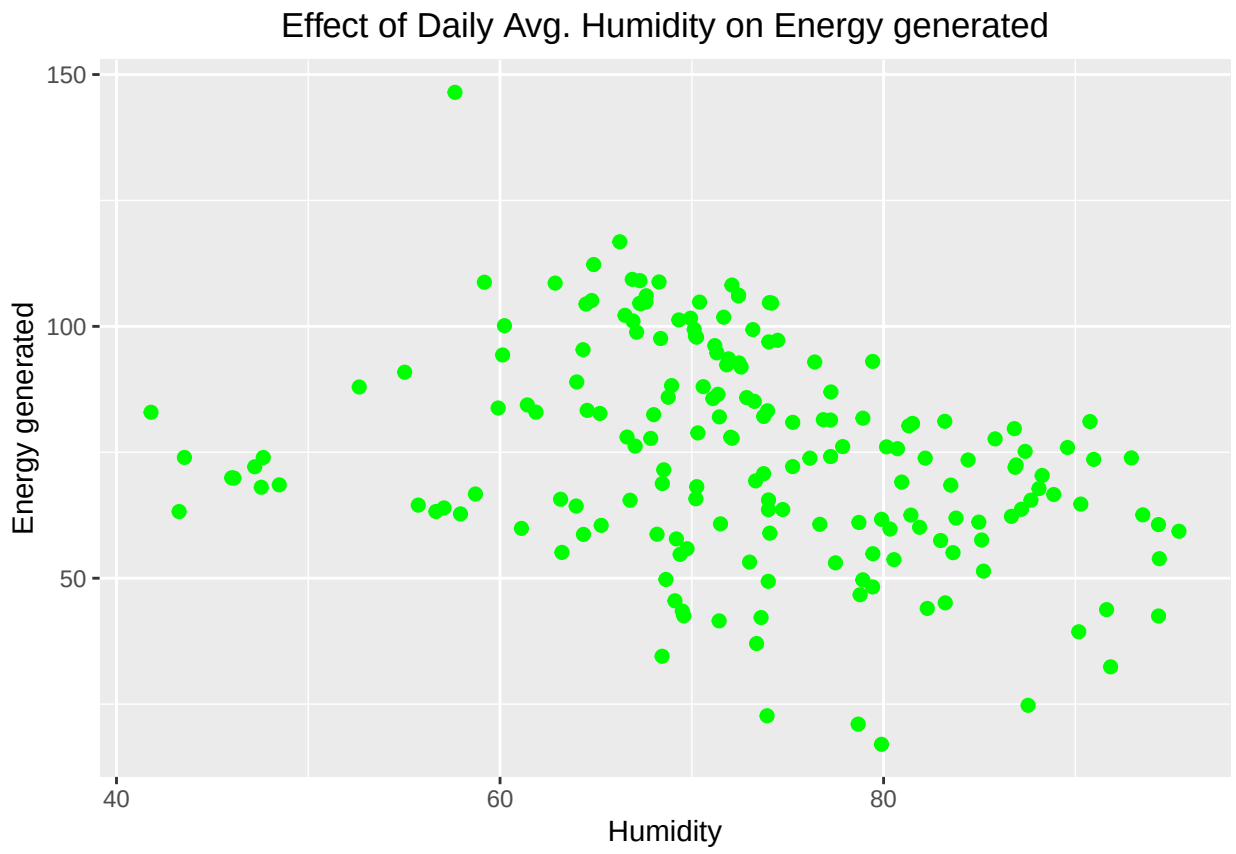

```
ggplot(by_day)+
  geom_point(mapping = aes(x = DNI, y = `energy_(Wh)`), colour = "darkblue", size = 2)+
  ylab("Energy generated")+
  xlab("Direct Normal Irradiance")+
  ggtitle("Effect of Daily Avg. Direct Normal Irradiance on Energy generated")+
  theme(plot.title = element_text(hjust = 0.5))
```



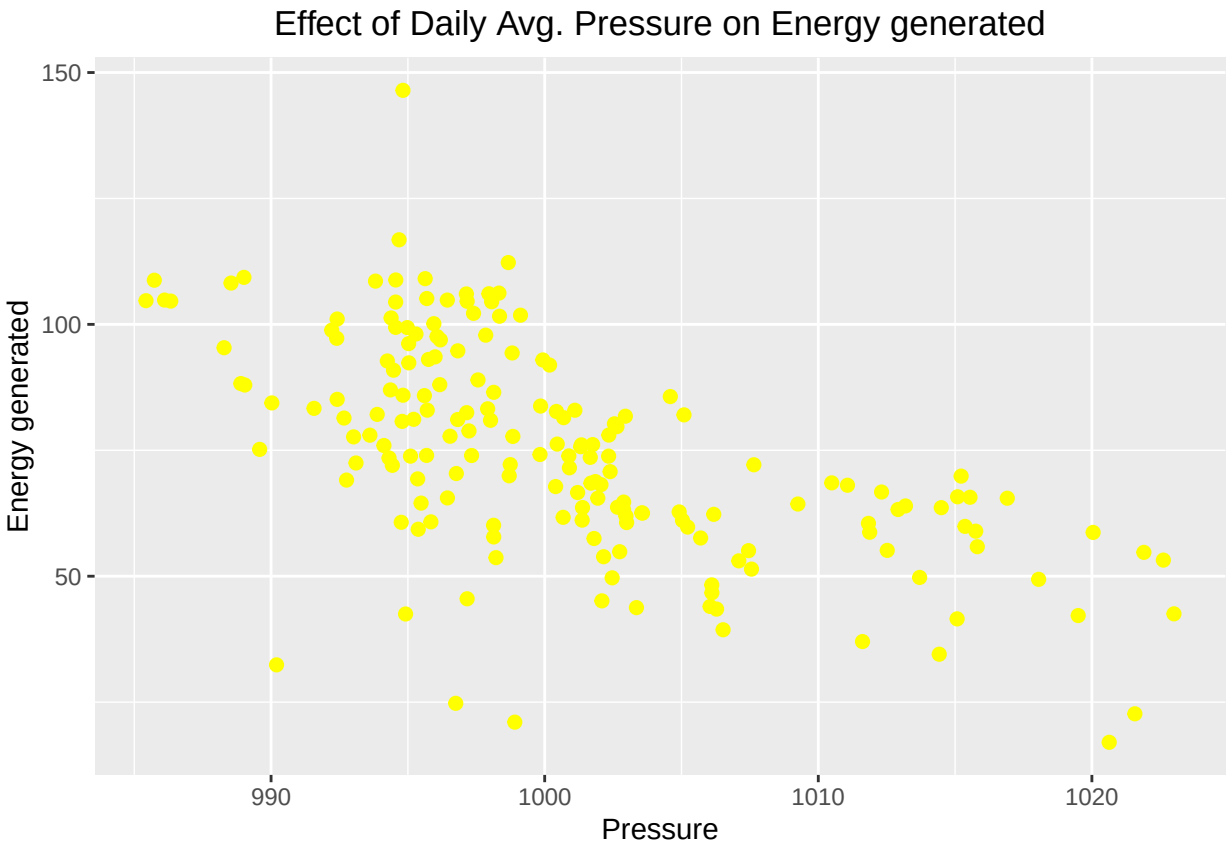

```
ggplot(by_day)+
  geom_point(mapping = aes(x = GHI, y = `energy_(Wh)`), colour = "orange", size = 2)+
  ylab("Energy generated")+
  xlab("Global Horizontal Irradiance")+
  ggtitle("Effect of Daily Avg. Global Horizontal Irradiance on Energy generated")+
  theme(plot.title = element_text(hjust = 0.5))
```



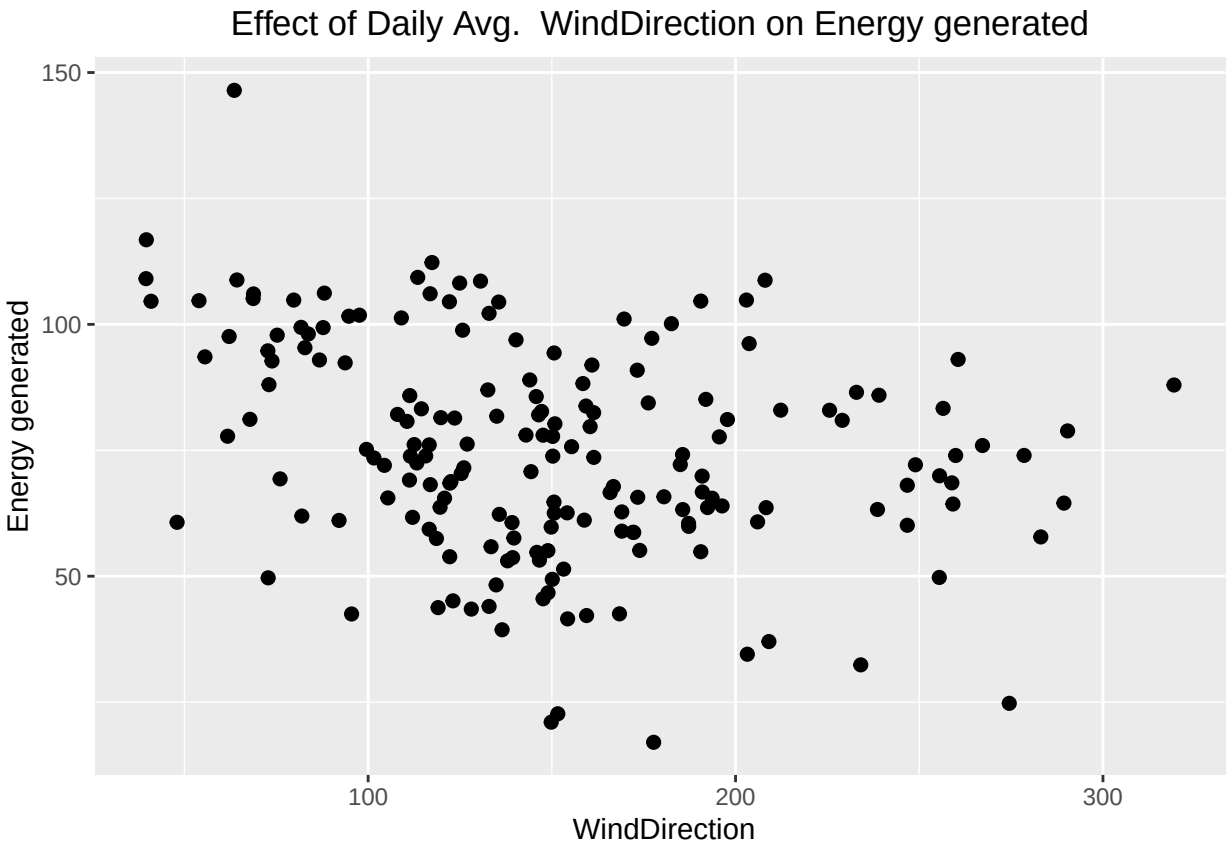

```
ggplot(by_day)+
  geom_point(mapping = aes(x = Humidity, y = `energy_(Wh)`), colour = "green", size = 2)+
  ylab("Energy generated")+
  xlab("Humidity")+
  ggtitle("Effect of Daily Avg. Humidity on Energy generated")+
  theme(plot.title = element_text(hjust = 0.5))
```



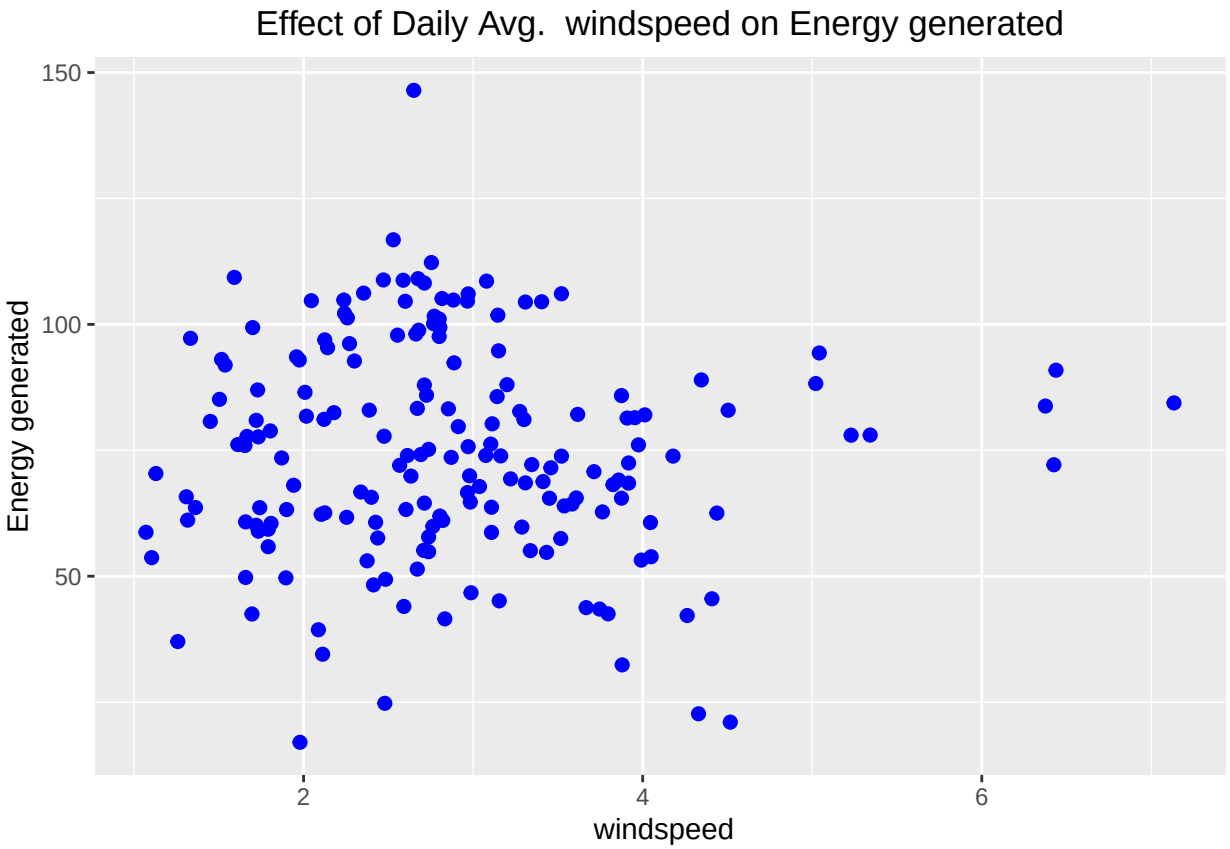

```
ggplot(by_day)+
  geom_point(mapping = aes(x = Pressure, y = `energy_(Wh)`), colour = "yellow", size = 2)+
  ylab("Energy generated")+
  xlab("Pressure")+
  ggtitle("Effect of Daily Avg. Pressure on Energy generated")+
  theme(plot.title = element_text(hjust = 0.5))
```




```
ggplot(by_day)+
  geom_point(mapping = aes(x =  WindDirection, y = `energy_(Wh)`), colour = "black", size = 2)+
  ylab("Energy generated")+
  xlab(" WindDirection")+
  ggtitle("Effect of Daily Avg.  WindDirection on Energy generated")+
  theme(plot.title = element_text(hjust = 0.5))
```




```
ggplot(by_day)+
  geom_point(mapping = aes(x =  windspeed, y = `energy_(Wh)`), colour = "blue", size = 2)+
  ylab("Energy generated")+
  xlab(" windspeed")+
  ggtitle("Effect of Daily Avg.  windspeed on Energy generated")+
  theme(plot.title = element_text(hjust = 0.5))
```



Buliding the Model

Building a model using simple linear regression and multiple regression mode to analyse the relation between the weather variables and the PV generated energy by considering `energy_(Wh)` as the response variable and 'Ghi', 'SurfacePressure', 'AirTemp','RelativeHumidity','Windspeed','winddirection' as the predictor variables.

```
### splitting the data into train and test data
library(caTools)
```

```
## Warning: package 'caTools' was built under R version 4.0.3
```

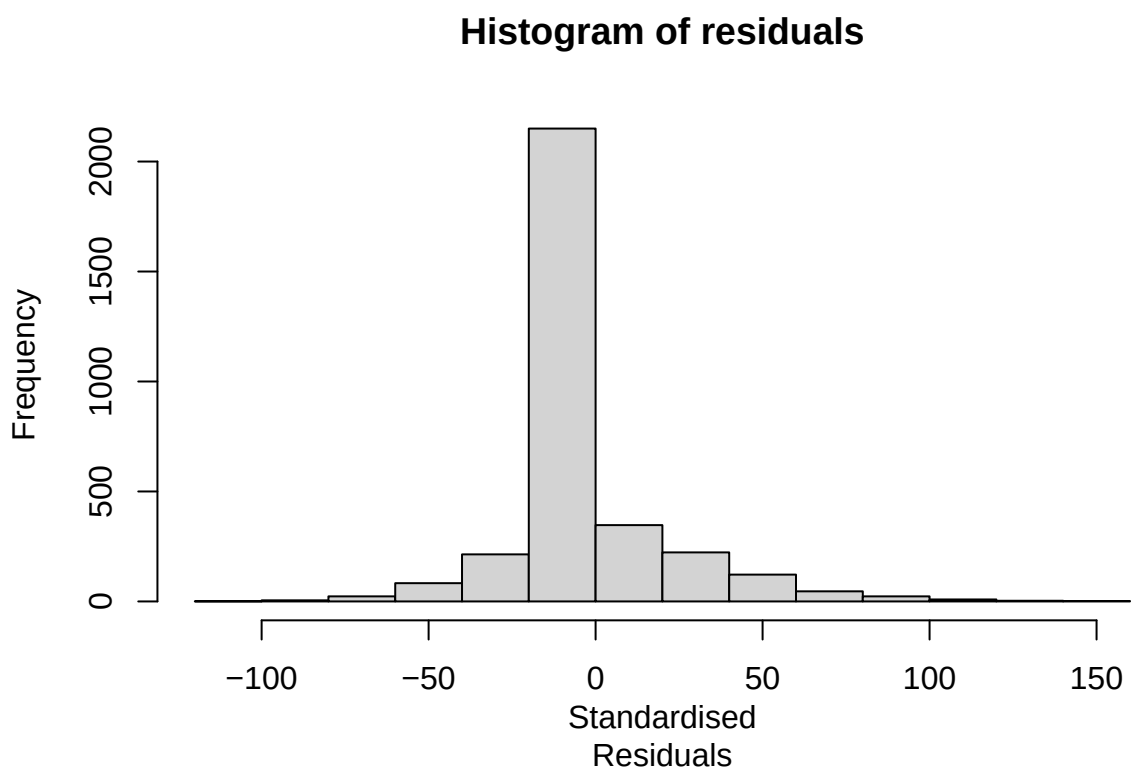
```
set.seed(10000)
sample = sample.split(solar_weather,SplitRatio = 0.80) # splits the data in the ratio mentioned in Spli
train1 =subset(solar_weather,sample ==TRUE) # creates a training dataset named train1 with rows which a
test1=subset(solar_weather, sample==FALSE)
#summary(train1)
#summary(test1)
solar_weather <- na.omit(solar_weather)
```

```
# Fitting simple linear regtression model
fit1<- lm(`energy_(Wh)`~Ghi, data = train1)
summary(fit1)
```

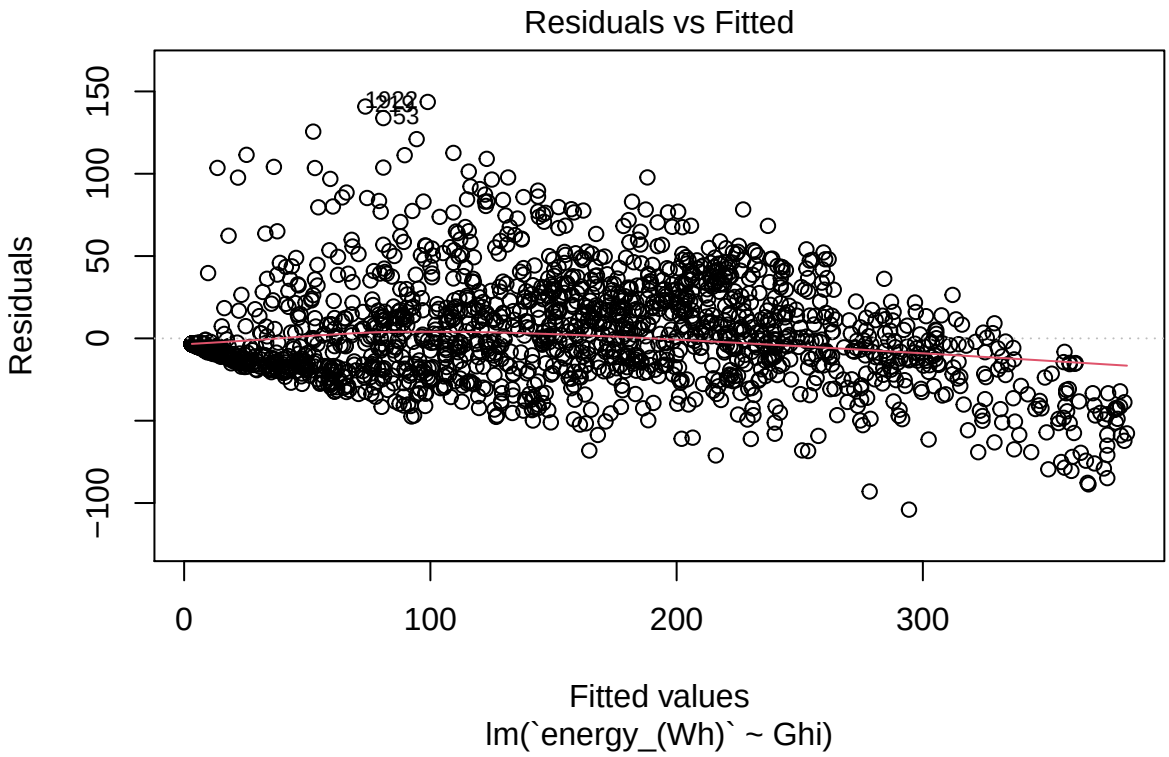
```
##
## Call:
## lm(formula = `energy_(Wh)` ~ Ghi, data = train1)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -104.007   -3.767   -3.251   -1.669  143.590
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.128410   0.513226   6.096 1.22e-09 ***
## Ghi         0.347117   0.001387 250.250  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.17 on 3249 degrees of freedom
## Multiple R-squared:  0.9507, Adjusted R-squared:  0.9507
## F-statistic: 6.263e+04 on 1 and 3249 DF,  p-value: < 2.2e-16
```

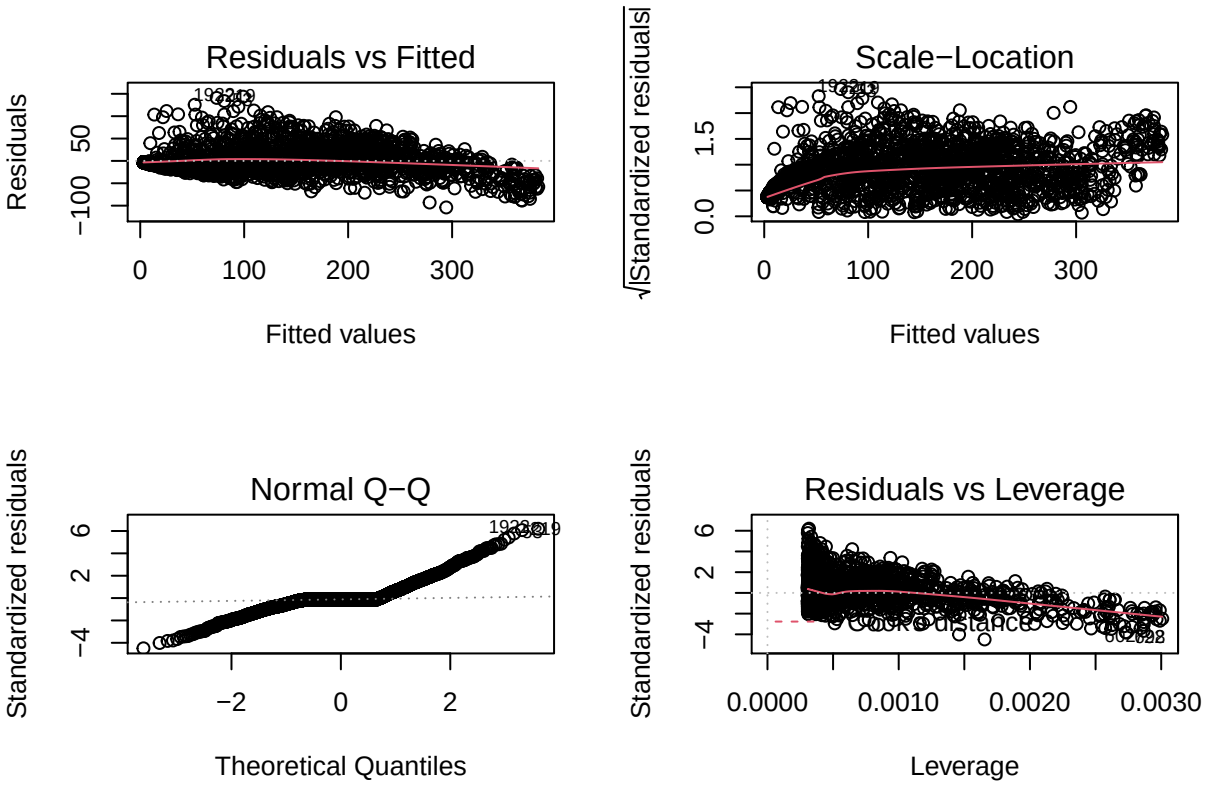
```
#Computing the accuracy of the model with HIgher Fstatictic
hist(resid(fit1),main='Histogram of residuals',xlab='Standardised
Residuals',ylab='Frequency')
```




```
plot(fit1, which =1)
```




```
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(fit1)
```




```
predictions<-predict(fit1, test1)
mse <- mean((test1$`energy_(Wh)` - predictions)^2)
print(mse)
```

```
## [1] 408.8205
```

```
sigma(fit1)/mean(test1$`energy_(Wh)`)
```

```
## [1] 0.4372723
```

```
test1$predicted<- predict(fit1,test1)
actuals_preds <- data.frame(test1$`energy_(Wh)`,test1$predicted)
names(actuals_preds)<- c("Energy_out","predicted")
correlation_accuracy <- cor(actuals_preds)
correlation_accuracy
```

```
##            Energy_out predicted
## Energy_out  1.0000000 0.9743054
## predicted   0.9743054 1.0000000
```

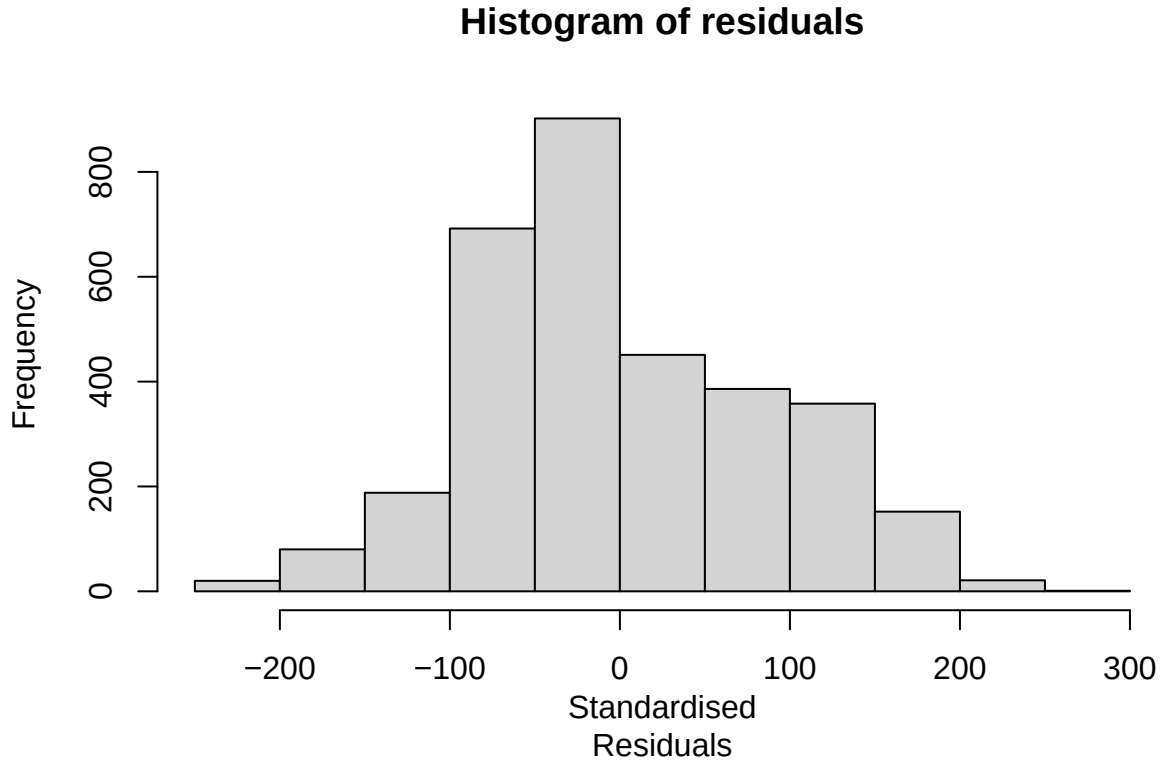
```
#head(actuals_preds)
```

```
fit2<- lm(`energy_(Wh)`~RelativeHumidity, data = train1)
summary(fit2)
```

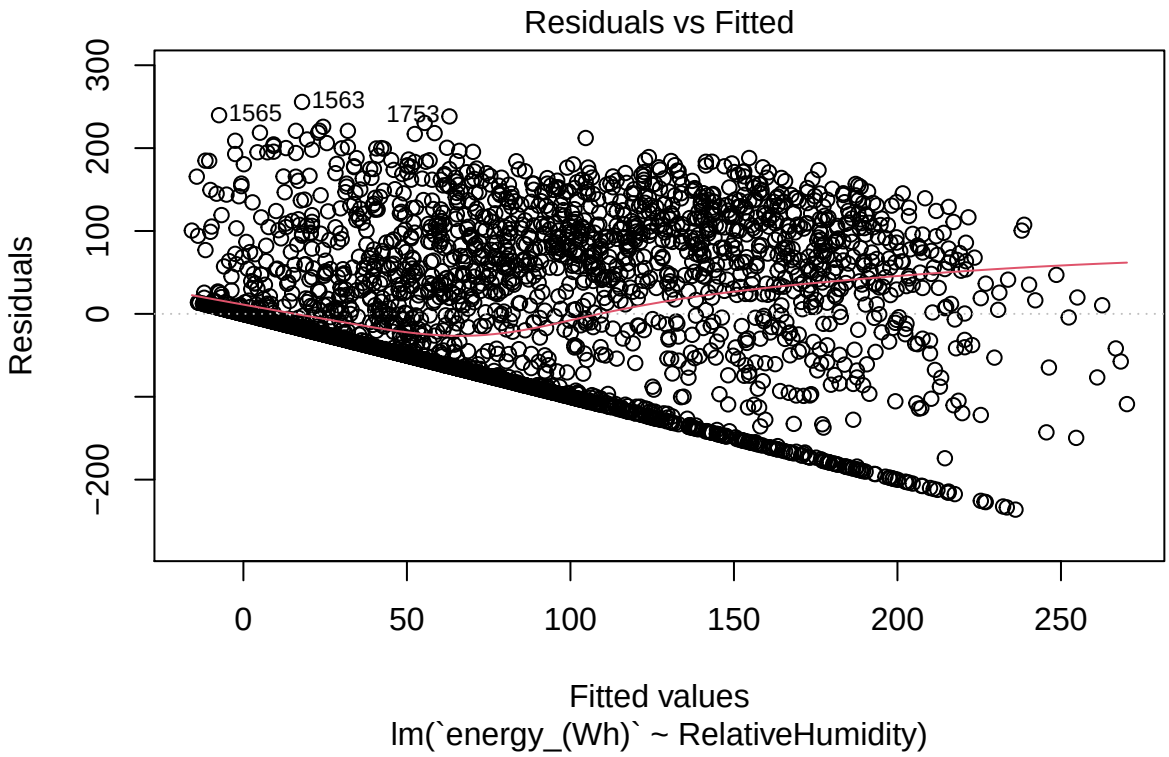
```
##
## Call:
## lm(formula = `energy_(Wh)` ~ RelativeHumidity, data = train1)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -236.19  -59.32  -14.54   63.52  255.69
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      357.66154    7.18986   49.74   <2e-16 ***
## RelativeHumidity  -3.78731    0.09645  -39.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 85.9 on 3249 degrees of freedom
## Multiple R-squared:  0.3219, Adjusted R-squared:  0.3216
## F-statistic:  1542 on 1 and 3249 DF,  p-value: < 2.2e-16
```

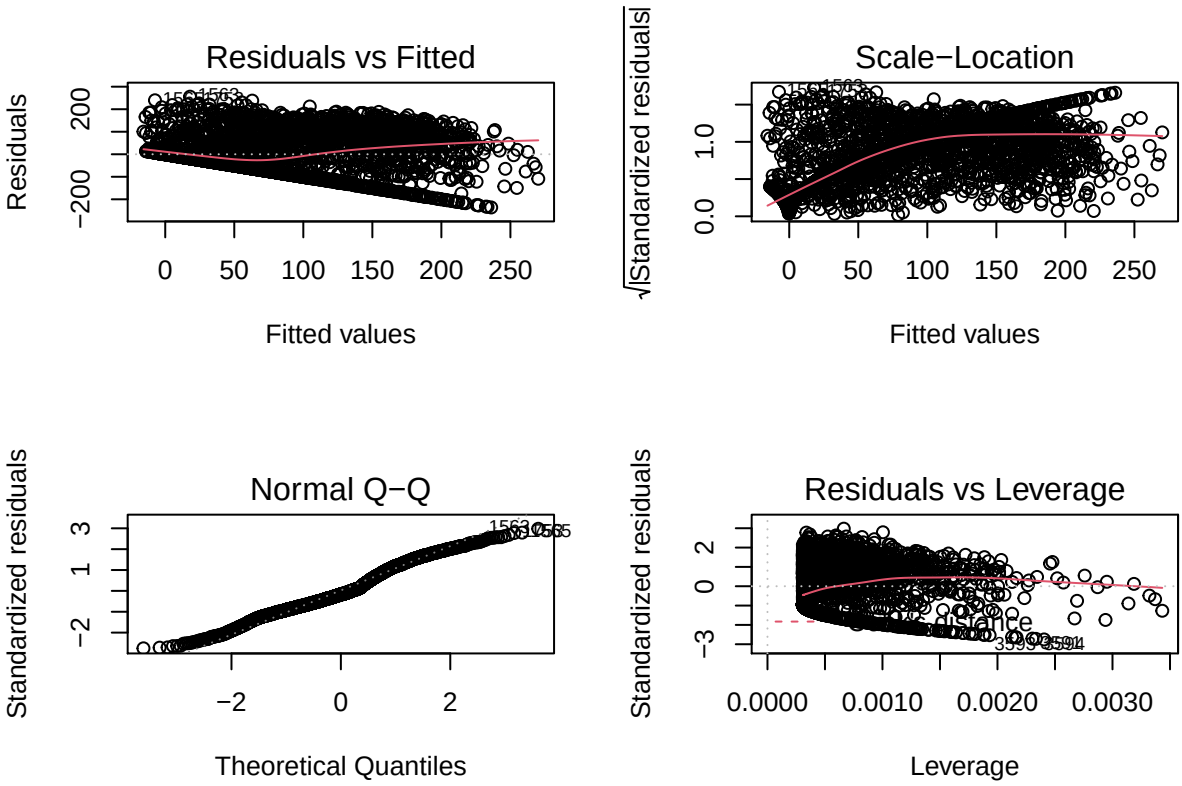
```
#Computing the accuracy of the model with HIgher Fstatictic
hist(resid(fit2),main='Histogram of residuals',xlab='Standardised
Residuals',ylab='Frequency')
```




```
plot(fit2, which =1)
```




```
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(fit2)
```




```
predictions<-predict(fit2, test1)
mse <- mean((test1$`energy_(Wh)` - predictions)^2)
print(mse)
```

```
## [1] 6467.665
```

```
sigma(fit2)/mean(test1$`energy_(Wh)`)
```

```
## [1] 1.621418
```

```
test1$predicted<- predict(fit2,test1)
actuals_preds <- data.frame(test1$`energy_(Wh)`,test1$predicted)
names(actuals_preds)<- c("Energy_out","predicted")
correlation_accuracy <- cor(actuals_preds)
correlation_accuracy
```

```
##            Energy_out predicted
## Energy_out  1.0000000 0.5764313
## predicted   0.5764313 1.0000000
```

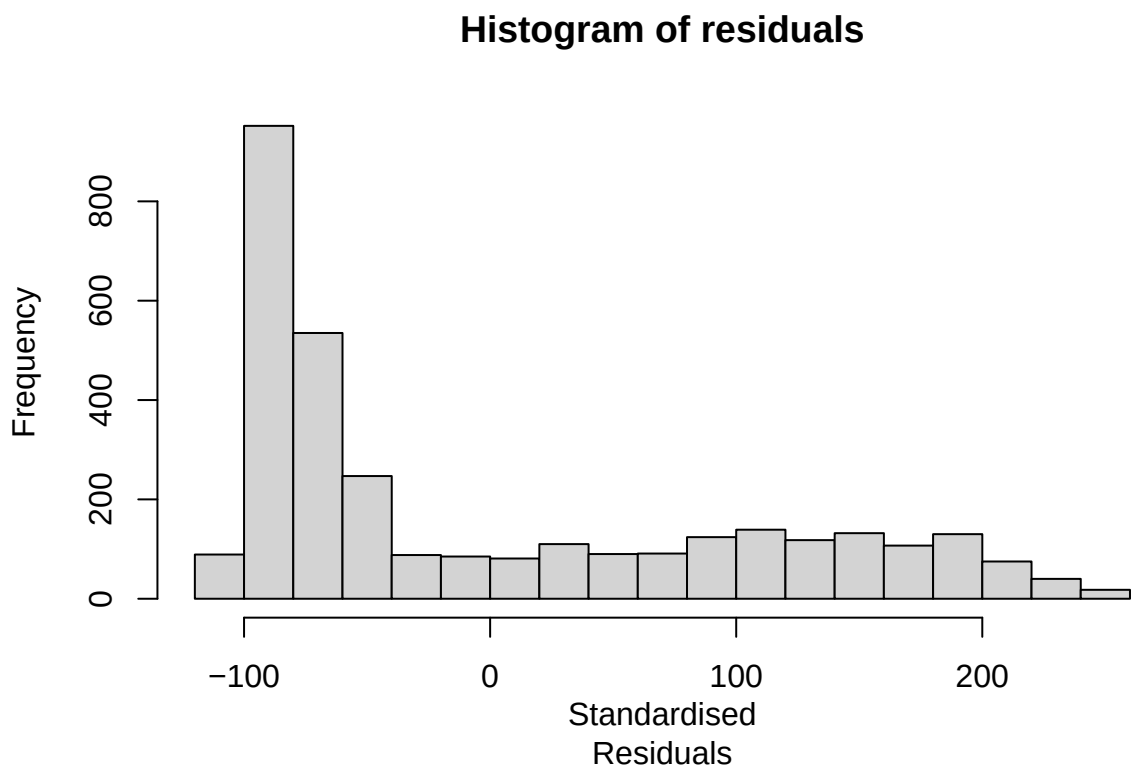
```
head(actuals_preds)
```

```
##    Energy_out  predicted
## 1 334.2183710 194.049835
## 2 135.3936369 175.870756
## 3  53.0971803 152.389447
## 4  -0.1112348   4.305704
## 5  -0.1121025  13.773974
## 6   9.6602855  28.544475
```

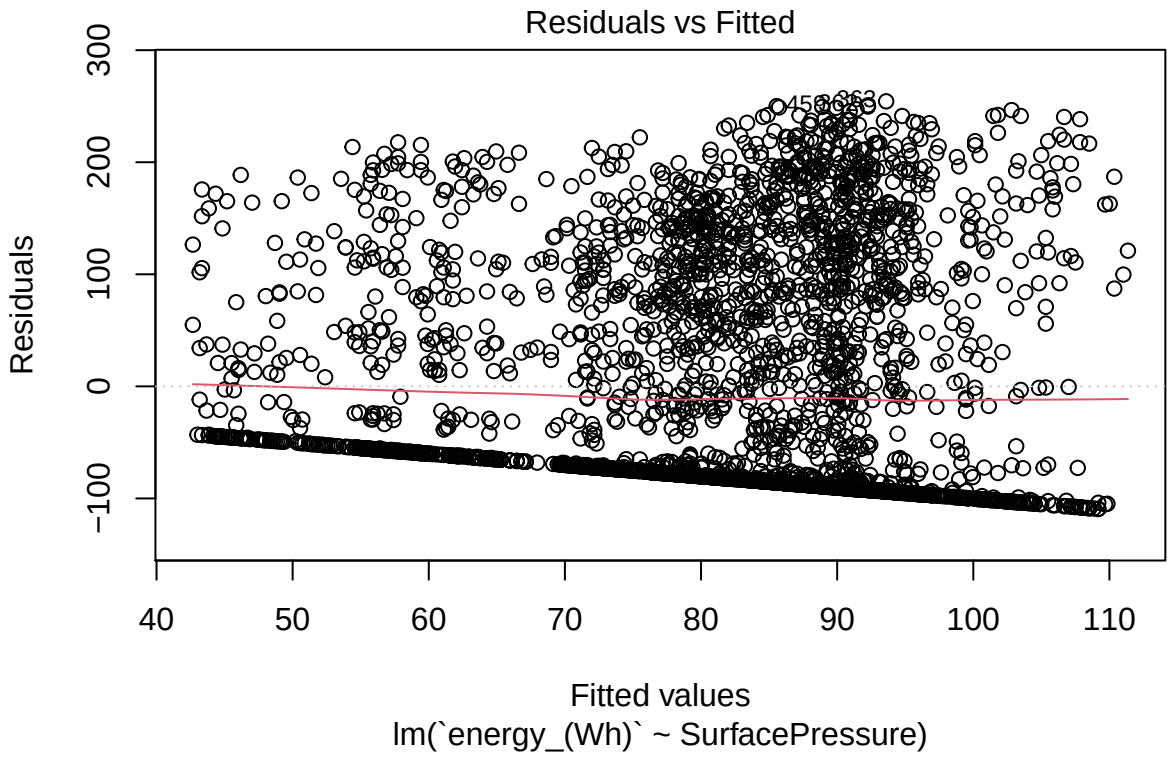
```r
fit3<- lm(`energy_(Wh)`~SurfacePressure, data = train1)
summary(fit3)
```

```
##
## Call:
## lm(formula = `energy_(Wh)` ~ SurfacePressure, data = train1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -109.42  -84.19  -57.71   88.92  254.34
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1758.6507   224.6798   7.827 6.69e-15 ***
## SurfacePressure   -1.6754     0.2245  -7.464 1.07e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103.4 on 3249 degrees of freedom
## Multiple R-squared:  0.01686,    Adjusted R-squared:  0.01656
## F-statistic: 55.72 on 1 and 3249 DF,  p-value: 1.068e-13
```

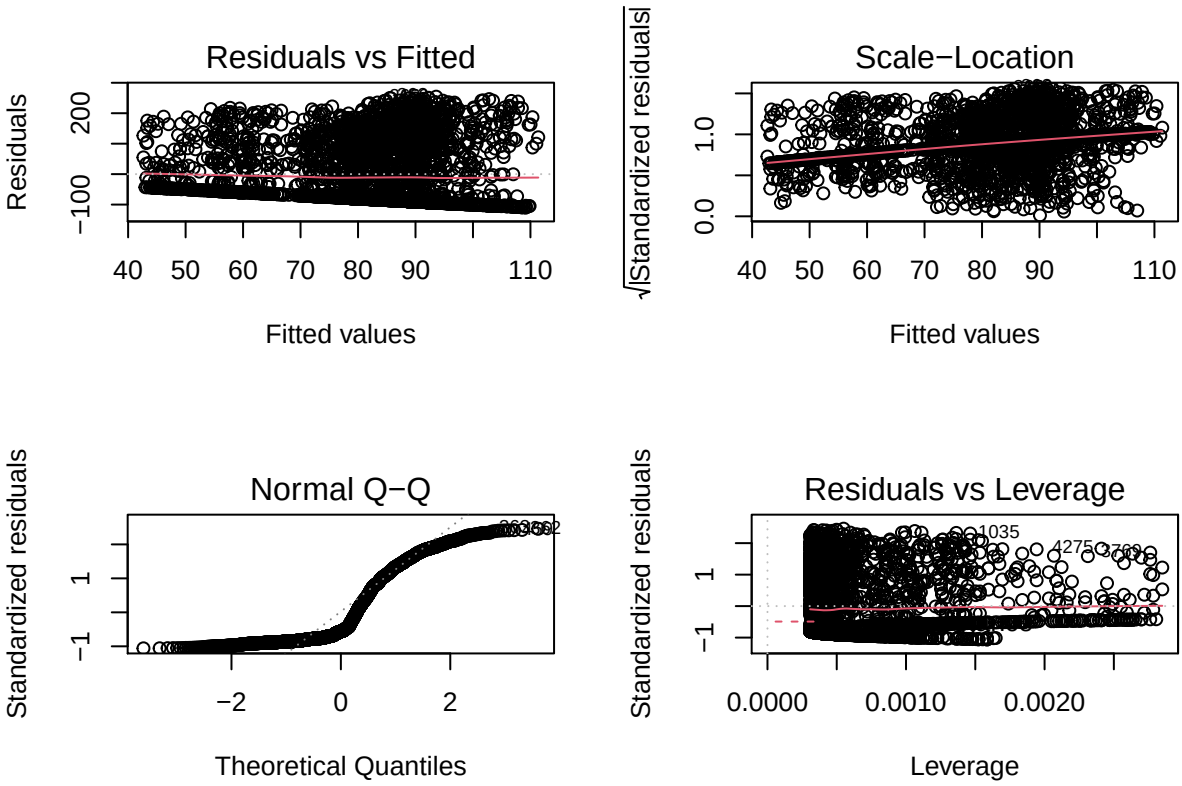
```r
#Computing the accuracy of the model with HIgher Fstatictic
hist(resid(fit3),main='Histogram of residuals',xlab='Standardised
Residuals',ylab='Frequency')
```



```r
plot(fit3, which =1)
```




```r
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(fit3)
```




```
predictions<-predict(fit3, test1)
mse <- mean((test1$`energy_(Wh)` - predictions)^2)
print(mse)
```

```
## [1] 8609.909
```

```
sigma(fit3)/mean(test1$`energy_(Wh)`)
```

```
## [1] 1.95228
```

```
test1$predicted<- predict(fit3,test1)
actuals_preds <- data.frame(test1$`energy_(Wh)`,test1$predicted)
names(actuals_preds)<- c("Energy_out","predicted")
correlation_accuracy <- cor(actuals_preds)
correlation_accuracy
```

```
##            Energy_out predicted
## Energy_out  1.0000000 0.1781231
## predicted   0.1781231 1.0000000
```

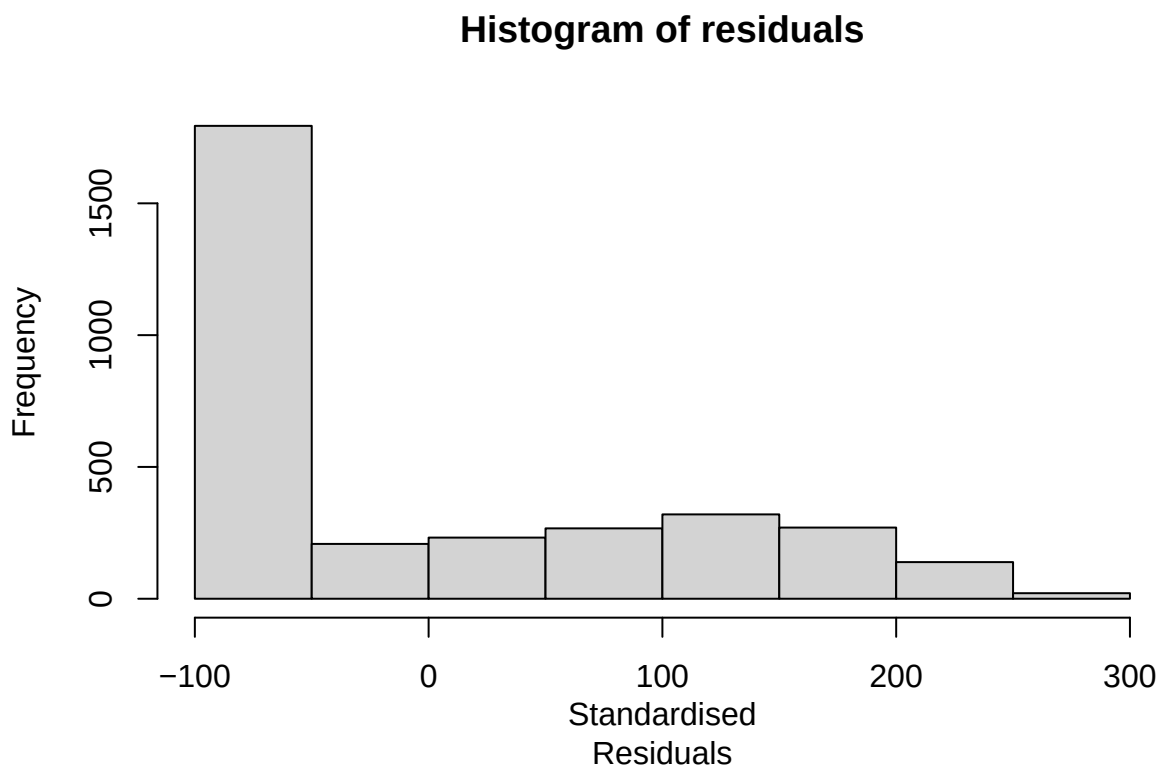
```
head(actuals_preds)
```

```
##    Energy_out predicted
## 1 334.2183710  93.26646
## 2 135.3936369  94.27172
## 3  53.0971803  93.09891
## 4  -0.1112348  92.09365
## 5  -0.1121025  91.75856
## 6   9.6602855  90.92085
```

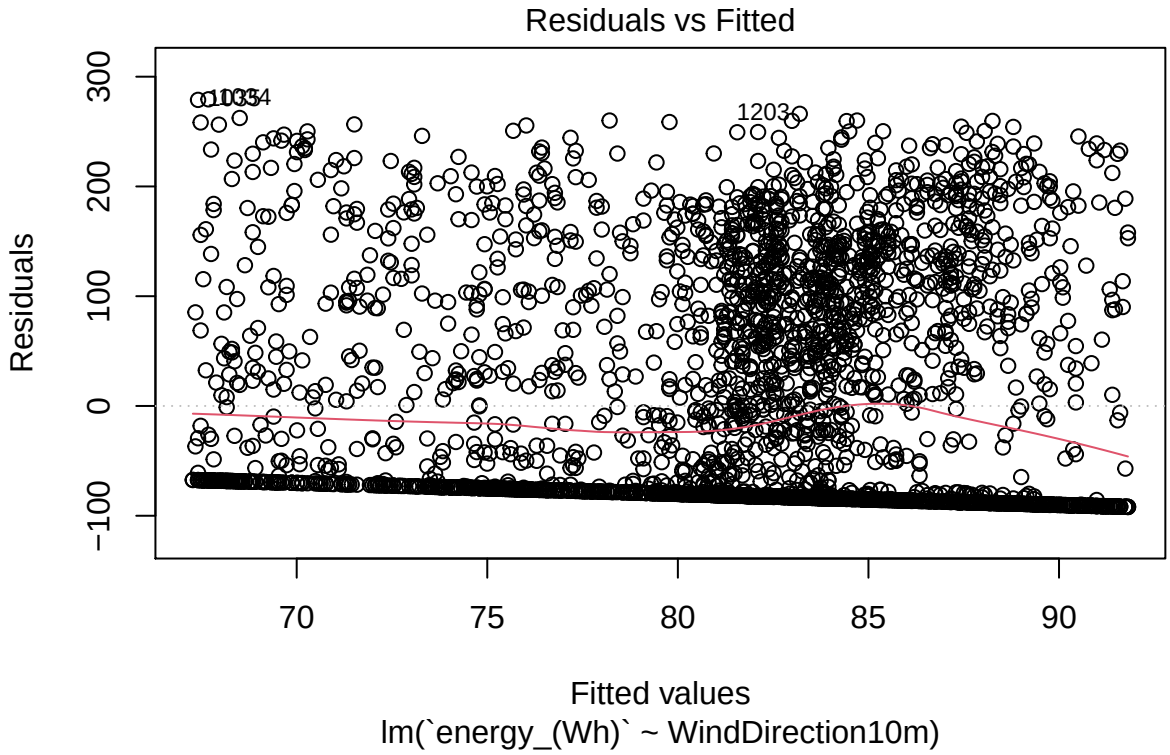
```r
fit4<- lm(`energy_(Wh)`~WindDirection10m, data = train1)
summary(fit4)
```

```
##
## Call:
## lm(formula = `energy_(Wh)` ~ WindDirection10m, data = train1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -92.03 -82.36 -69.32  90.18 279.37
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      91.87687    3.74725  24.518  < 2e-16 ***
## WindDirection10m -0.06834    0.02174  -3.143  0.00169 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 104.2 on 3249 degrees of freedom
## Multiple R-squared:  0.003031,   Adjusted R-squared:  0.002724
## F-statistic: 9.878 on 1 and 3249 DF,  p-value: 0.001688
```

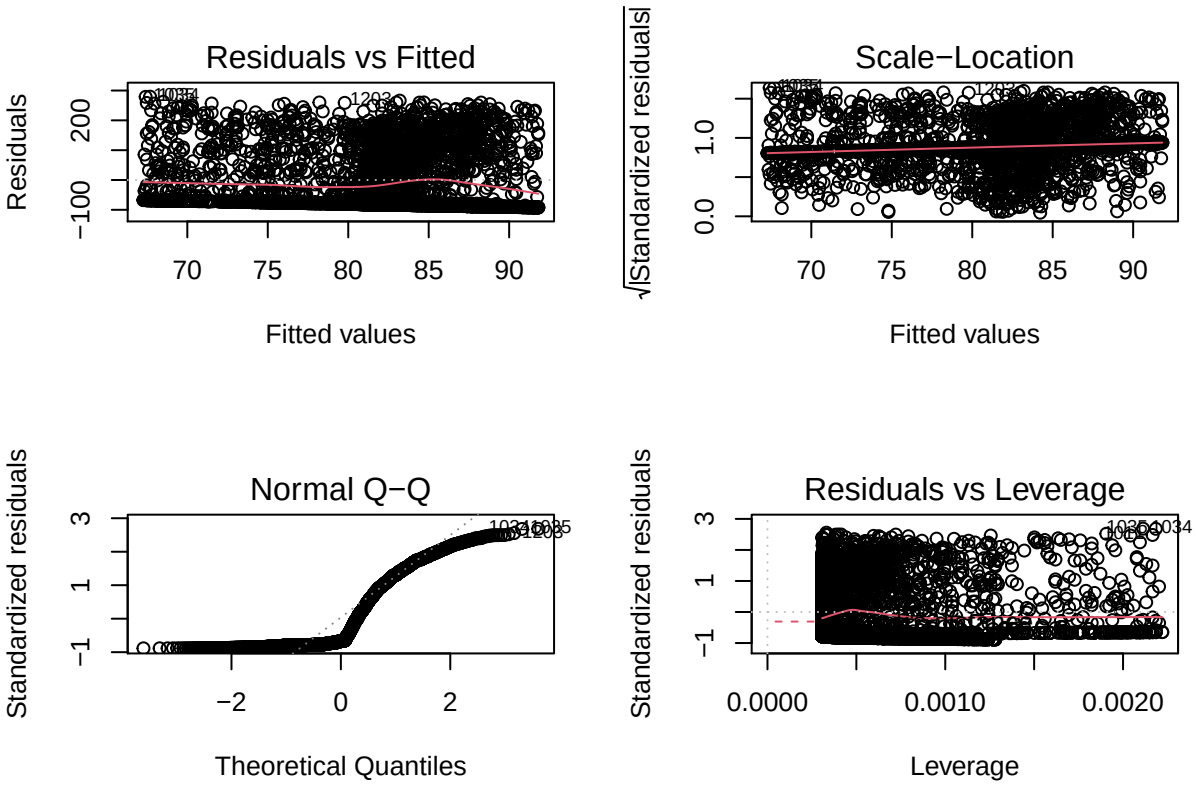
```r
#Computing the accuracy of the model with HIgher Fstatictic
hist(resid(fit4),main='Histogram of residuals',xlab='Standardised
Residuals',ylab='Frequency')
```



```
plot(fit4, which =1)
```




```
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(fit4)
```




```
predictions<-predict(fit4, test1)
mse <- mean((test1$`energy_(Wh)` - predictions)^2)
print(mse)
```

```
## [1] 8605.434
```

```
sigma(fit4)/mean(test1$`energy_(Wh)`)
```

```
## [1] 1.965962
```

```
test1$predicted<- predict(fit4,test1)
actuals_preds <- data.frame(test1$`energy_(Wh)`,test1$predicted)
names(actuals_preds)<- c("Energy_out","predicted")
correlation_accuracy <- cor(actuals_preds)
correlation_accuracy
```

```
##            Energy_out predicted
## Energy_out  1.0000000 0.2131693
## predicted   0.2131693 1.0000000
```

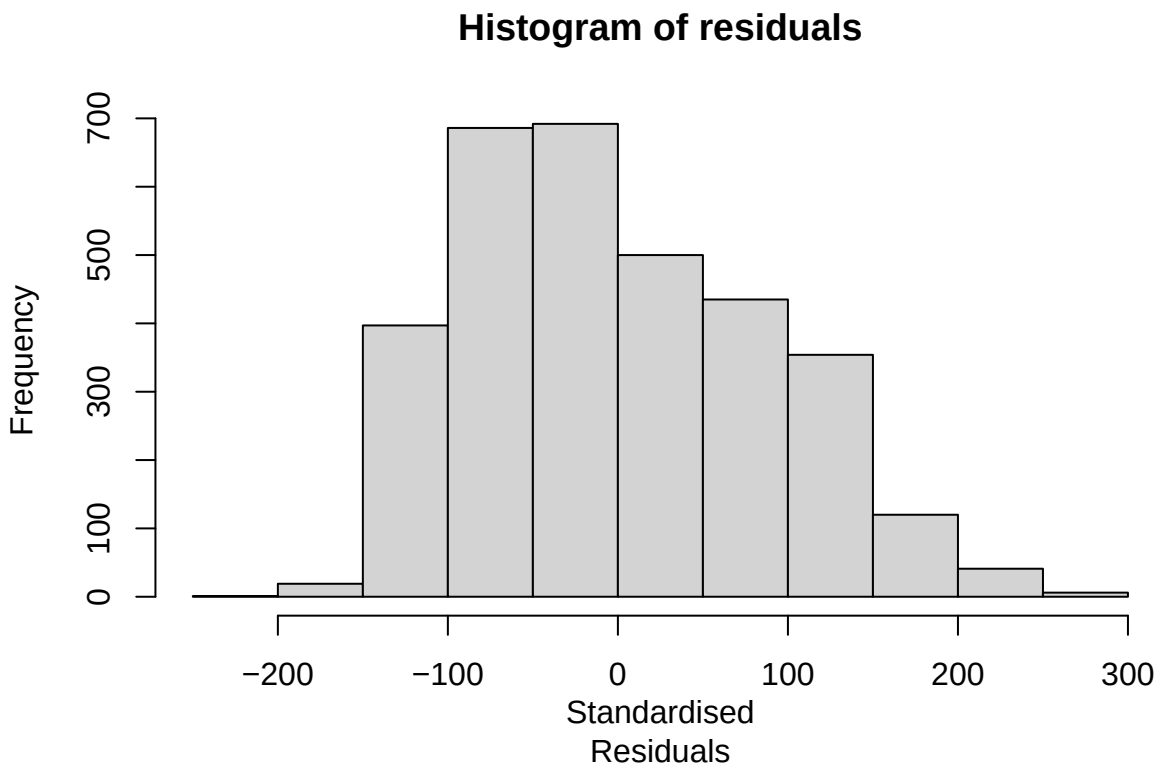
```
head(actuals_preds)
```

```
##    Energy_out predicted
## 1 334.2183710  87.29824
## 2 135.3936369  87.29824
## 3  53.0971803  87.43492
## 4  -0.1112348  87.91328
## 5  -0.1121025  87.16157
## 6   9.6602855  87.16157
```

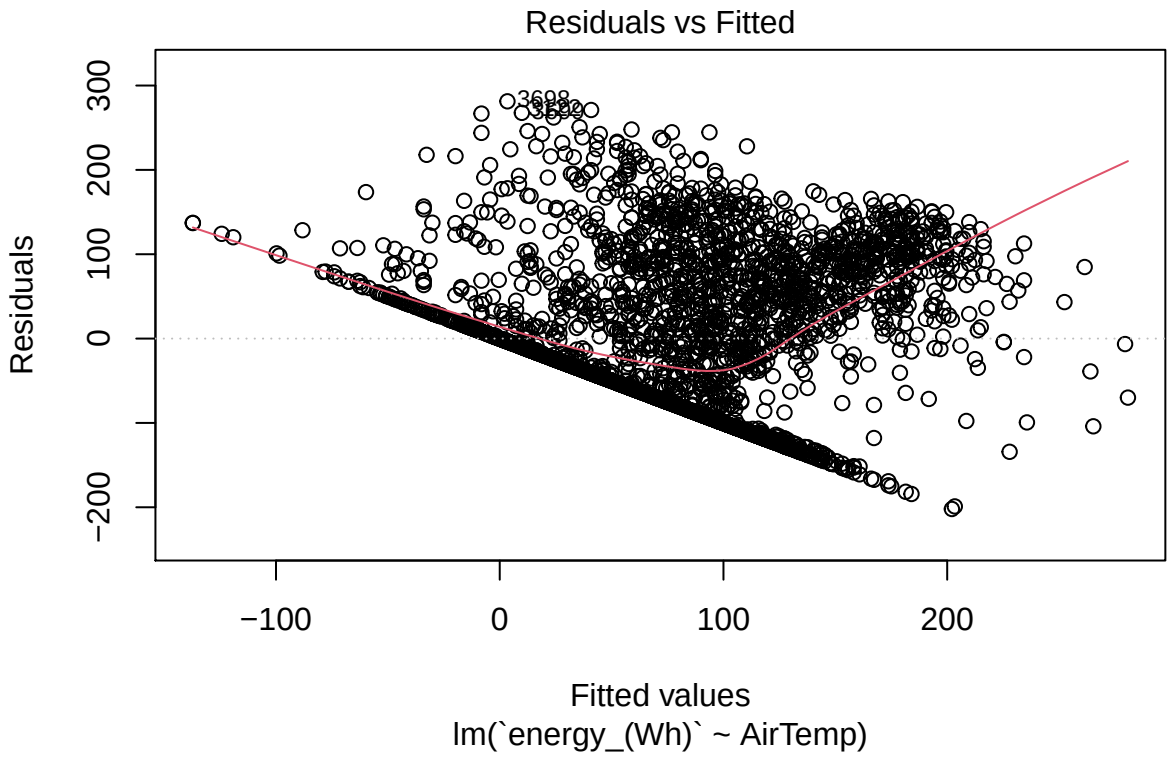
```r
fit6<- lm(`energy_(Wh)`~AirTemp, data = train1)
summary(fit6)
```

```
##
## Call:
## lm(formula = `energy_(Wh)` ~ AirTemp, data = train1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -202.04  -70.73  -12.59   67.26  281.26
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -191.3489     7.5653  -25.29   <2e-16 ***
## AirTemp       12.9001     0.3501   36.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 87.61 on 3249 degrees of freedom
## Multiple R-squared:  0.2947, Adjusted R-squared:  0.2945
## F-statistic:  1358 on 1 and 3249 DF,  p-value: < 2.2e-16
```

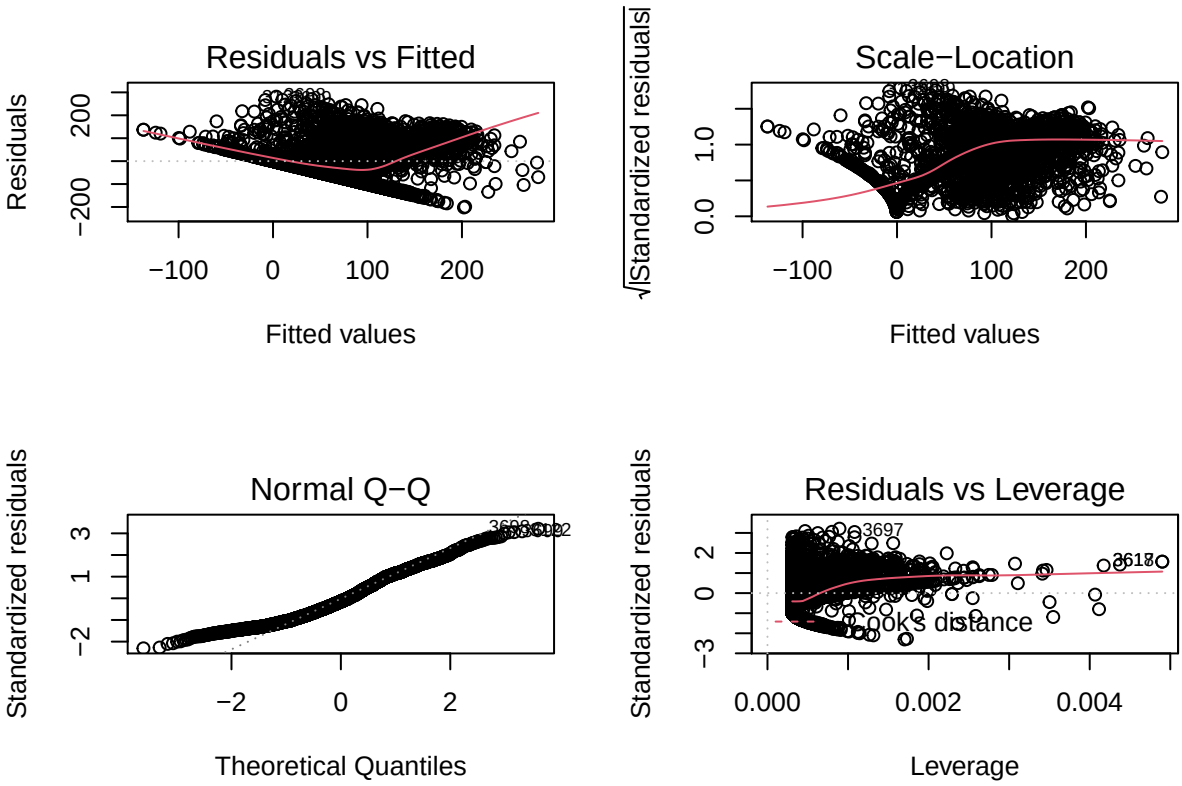
```r
#Computing the accuracy of the model with HIgher Fstatictic
hist(resid(fit6),main='Histogram of residuals',xlab='Standardised
Residuals',ylab='Frequency')
```



```r
plot(fit6, which =1)
```




```r
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(fit6)
```




```
predictions<-predict(fit6, test1)
mse <- mean((test1$`energy_(Wh)` - predictions)^2)
print(mse)
```

```
## [1] 5772.105
```

```
sigma(fit6)/mean(test1$`energy_(Wh)`)
```

```
## [1] 1.653552
```

```
test1$predicted<- predict(fit6,test1)
actuals_preds <- data.frame(test1$`energy_(Wh)`,test1$predicted)
names(actuals_preds)<- c("Energy_out","predicted")
correlation_accuracy <- cor(actuals_preds)
correlation_accuracy
```

```
##            Energy_out predicted
## Energy_out   1.000000  0.609013
## predicted    0.609013  1.000000
```

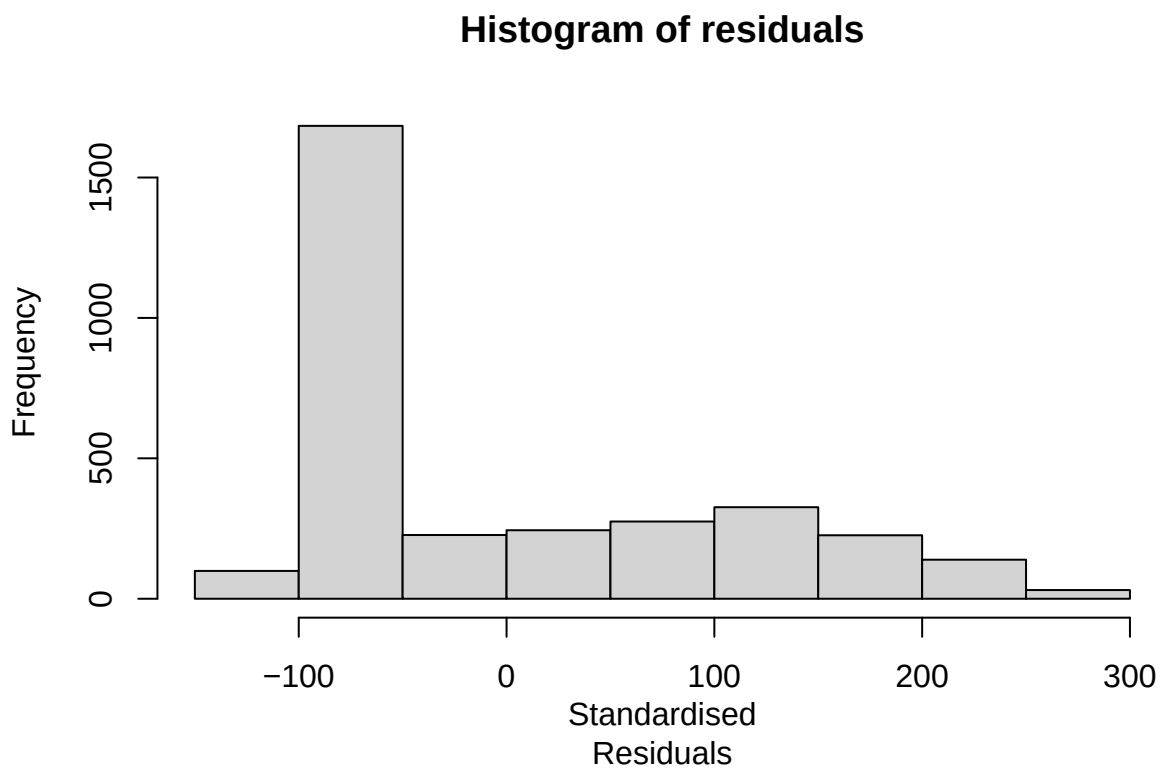
```
head(actuals_preds)
```

```
##    Energy_out predicted
## 1 334.2183710 181.46310
## 2 135.3936369 162.11300
## 3  53.0971803 142.76289
## 4  -0.1112348  57.62242
## 5  -0.1121025  55.04241
## 6   9.6602855  69.23249
```

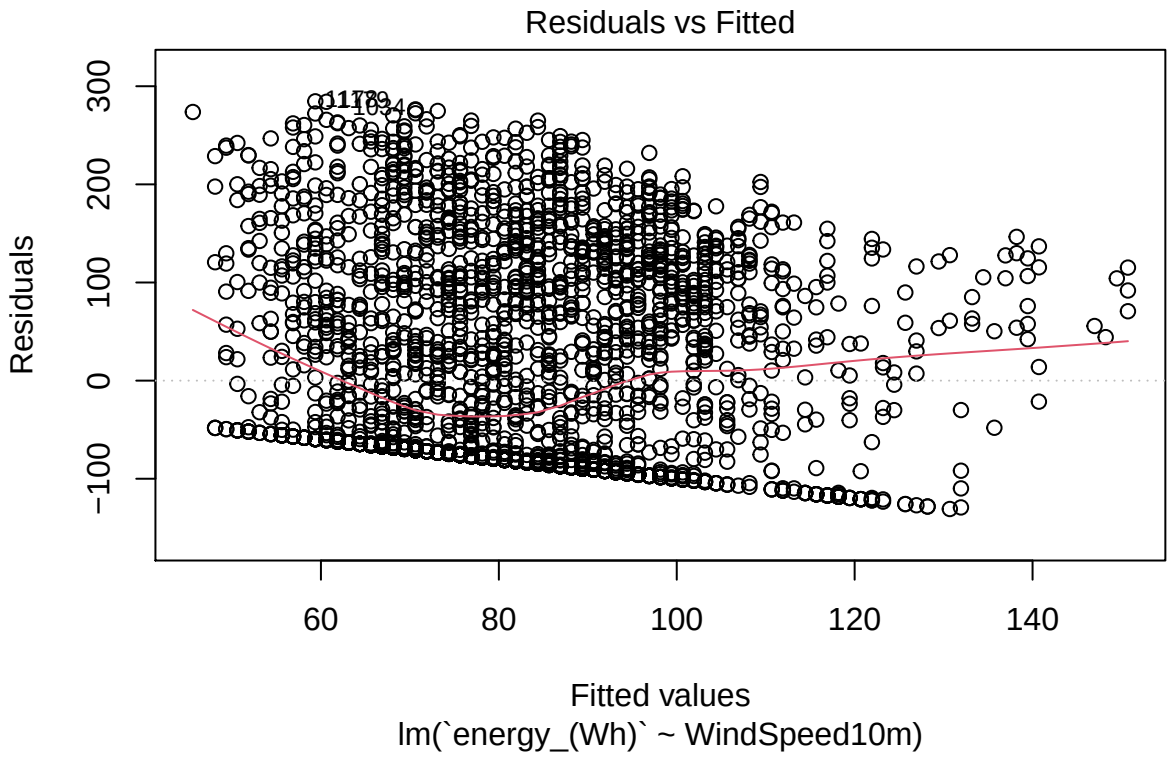
```r
fit5<- lm(`energy_(Wh)`~WindSpeed10m, data = train1)
summary(fit5)
```

```
##
## Call:
## lm(formula = `energy_(Wh)` ~ WindSpeed10m, data = train1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -130.83  -78.28  -63.27   85.18  284.61
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    45.600      4.473  10.194   <2e-16 ***
## WindSpeed10m   12.515      1.423   8.797   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103.1 on 3249 degrees of freedom
## Multiple R-squared:  0.02326,    Adjusted R-squared:  0.02296
## F-statistic: 77.38 on 1 and 3249 DF,  p-value: < 2.2e-16
```

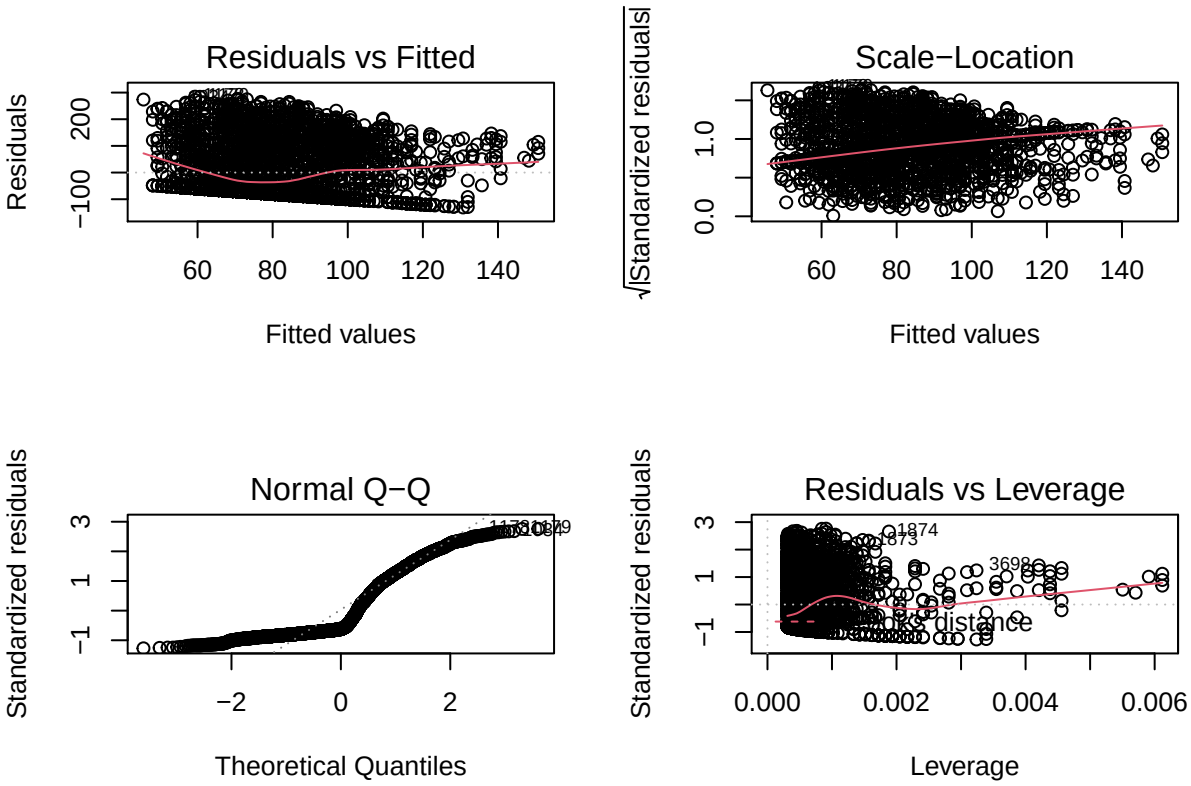
```r
#Computing the accuracy of the model with HIgher Fstatictic
hist(resid(fit5),main='Histogram of residuals',xlab='Standardised
Residuals',ylab='Frequency')
```



```r
plot(fit5, which =1)
```




```r
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(fit5)
```




```
predictions<-predict(fit5, test1)
mse <- mean((test1$`energy_(Wh)` - predictions)^2)
print(mse)
```

```
## [1] 8412.98
```

```
sigma(fit5)/mean(test1$`energy_(Wh)`)
```

```
## [1] 1.945911
```

```
test1$predicted<- predict(fit5,test1)
actuals_preds <- data.frame(test1$`energy_(Wh)`,test1$predicted)
names(actuals_preds)<- c("Energy_out","predicted")
correlation_accuracy <- cor(actuals_preds)
correlation_accuracy
```

```
##            Energy_out predicted
## Energy_out  1.0000000 0.2201767
## predicted   0.2201767 1.0000000
```

```
head(actuals_preds)
```

```
##    Energy_out predicted
## 1 334.2183710  79.39020
## 2 135.3936369  88.15069
## 3  53.0971803  86.89919
## 4  -0.1112348  70.62970
## 5  -0.1121025  70.62970
## 6   9.6602855  69.37820
```

```r
###Fitting Multi regression Model on the train data
fit7<- lm(`energy_(Wh)`~Ghi+AirTemp+RelativeHumidity+SurfacePressure+WindDirection10m+WindSpeed10m, data
summary(fit7)
```

```
##
## Call:
## lm(formula = `energy_(Wh)` ~ Ghi + AirTemp + RelativeHumidity +
##     SurfacePressure + WindDirection10m + WindSpeed10m, data = train1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -98.642  -9.685  -2.493   4.592 147.682
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -5.018e+02  6.571e+01  -7.637 2.90e-14 ***
## Ghi               3.343e-01  1.869e-03 178.838  < 2e-16 ***
## AirTemp           8.551e-01  1.473e-01   5.804 7.11e-09 ***
## RelativeHumidity -2.447e-01  3.130e-02  -7.818 7.19e-15 ***
## SurfacePressure   4.985e-01  6.319e-02   7.888 4.15e-15 ***
## WindDirection10m  1.665e-02  5.174e-03   3.218   0.0013 **
## WindSpeed10m      2.136e+00  3.191e-01   6.694 2.54e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.37 on 3244 degrees of freedom
## Multiple R-squared:  0.9541, Adjusted R-squared:  0.954
## F-statistic: 1.123e+04 on 6 and 3244 DF,  p-value: < 2.2e-16
```

```r
#Variable selection using stepwise aalysis
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.0.3
```

```
##
## Attaching package: 'MASS'
```

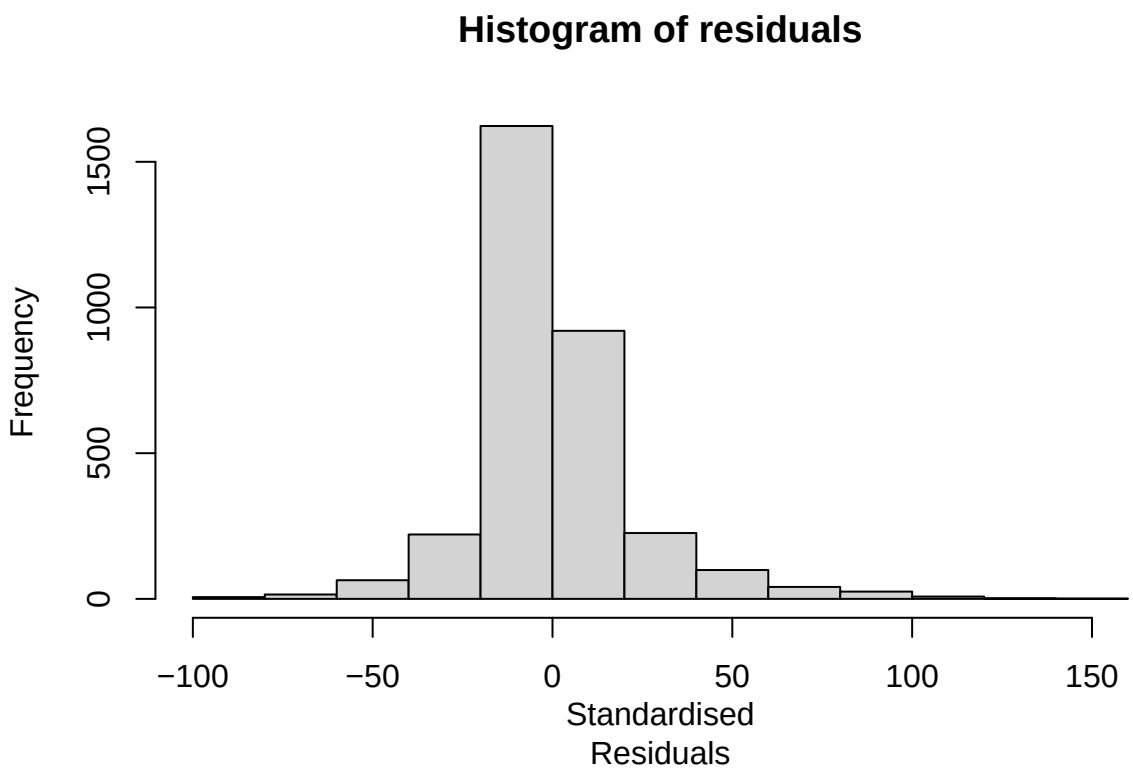
```
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
stepAIC(fit7, trace = FALSE)$anova
```

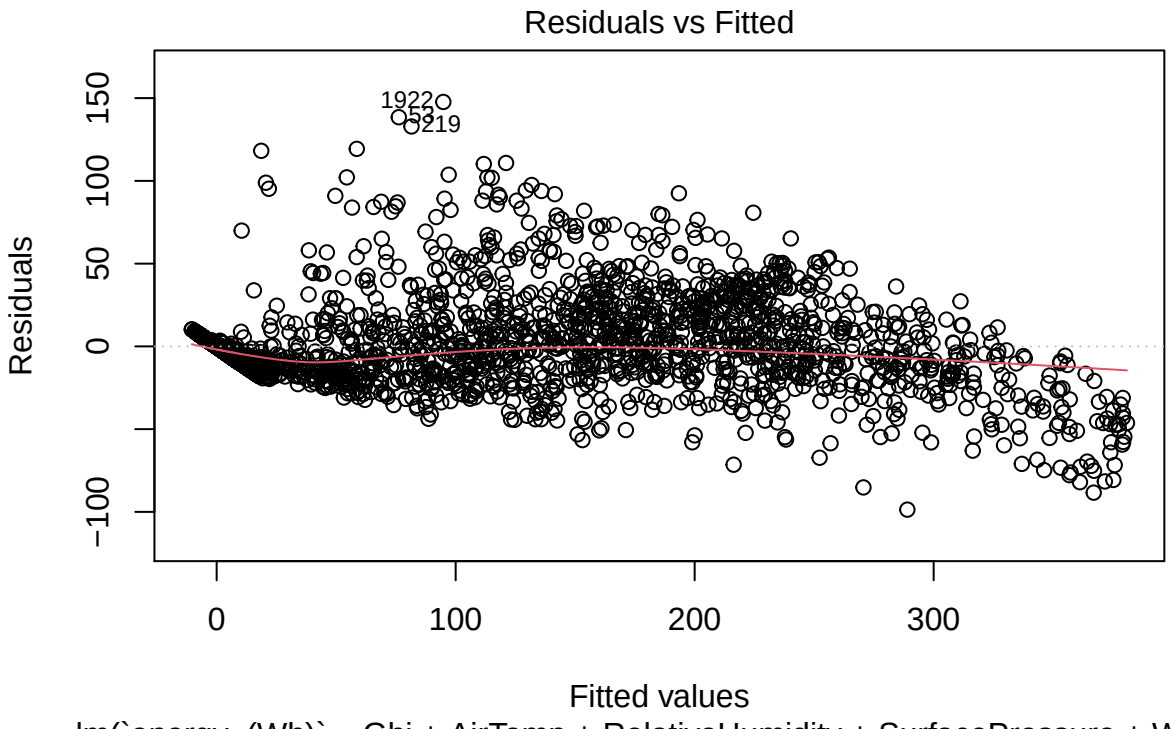
```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## `energy_(Wh)` ~ Ghi + AirTemp + RelativeHumidity + SurfacePressure +
##     WindDirection10m + WindSpeed10m
##
## Final Model:
## `energy_(Wh)` ~ Ghi + AirTemp + RelativeHumidity + SurfacePressure +
##     WindDirection10m + WindSpeed10m
##
##
##   Step Df Deviance Resid. Df Resid. Dev      AIC
```

```
## 1                         3244    1623550 20213.79
```

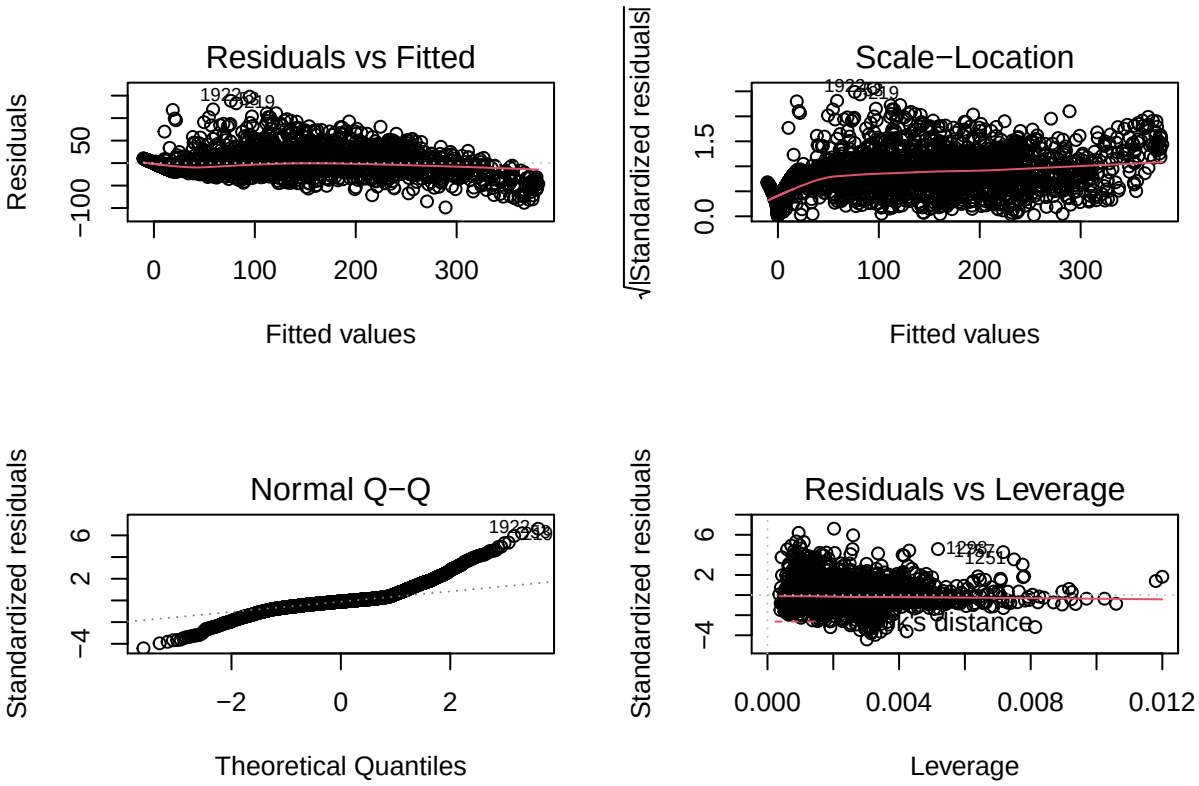
```
#Plot
hist(resid(fit7),main='Histogram of residuals',xlab='Standardised
Residuals',ylab='Frequency')
```




```
plot(fit7, which =1)
```




```
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(fit7)
```




```
#Computing Accuracy of the Model using the test data
predictions<-predict(fit7, test1)
mse <- mean((test1$`energy_(Wh)` - predictions)^2)
print(mse)
```

```
## [1] 422.332
```

```
sigma(fit7)/mean(test1$`energy_(Wh)`)
```

```
## [1] 0.4222566
```

```
test1$predicted<- predict(fit7,test1)
actuals_preds <- data.frame(test1$`energy_(Wh)`,test1$predicted)
names(actuals_preds)<- c("Energy_out","predicted")
correlation_accuracy <- cor(actuals_preds)
correlation_accuracy
```

```
##            Energy_out predicted
## Energy_out  1.0000000 0.9731785
## predicted   0.9731785 1.0000000
```

```
#head(actuals_preds)
```