A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Felipe Luiz Bauer Steudel |
| *Student Number* | 2024153 |
| *Module Title* | Strategic Thinking for Data Analysis |
| *Assessment Title* | CA2 |
| *Assessment Due Date* | 19-May-2024 |
| *Date of Submission* | 19-May-2024 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Contents

[Introduction 1](#_Toc167040602)

[Variable Identification 1](#_Toc167040603)

[Generation data 1](#_Toc167040604)

[Temperature and Solar Irradiation data 1](#_Toc167040605)

[The Employed Technique 2](#_Toc167040606)

[Exploratory Data Analysis 2](#_Toc167040607)

[Descriptive Statistics - Generation data 2](#_Toc167040608)

[Descriptive Statistics - Temperature and Solar Irradiation data 2](#_Toc167040609)

[Merging Dataframes 3](#_Toc167040610)

[Heatmap 3](#_Toc167040611)

[Matrix of Scatter Plots 4](#_Toc167040612)

[Data Processing Forecasting 4](#_Toc167040613)

[DC Power 5](#_Toc167040614)

[Histogram 5](#_Toc167040615)

[Scatter Plot and Mean DC Power 5](#_Toc167040616)

[Daily Yield 6](#_Toc167040617)

[Histogram 6](#_Toc167040618)

[Scatter Plot and Yield Mean 6](#_Toc167040619)

[Daily Yield and AC-DC Power 6](#_Toc167040620)

[Module Temperature 7](#_Toc167040621)

[Scatter Plot and Mean Temperature 7](#_Toc167040622)

[Application of Linear Regression 8](#_Toc167040623)

[Data Information 8](#_Toc167040624)

[Training and Test Sets to Linear Regression Model 8](#_Toc167040625)

[Coefficient of Determination ( R^2 ) 9](#_Toc167040626)

[MSE and MAE 9](#_Toc167040627)

[Findings 9](#_Toc167040628)

[Conclusions and Suggestions for the Future 10](#_Toc167040629)

[References 11](#_Toc167040630)

[GitHub Link 11](#_Toc167040631)

CA2 – Strategic Thinking for Data Analysis

# Introduction

This project's analysis of temperature sensor and photovoltaic power generating data from an Indian solar plant is its goal. The datasets are on Kaggle and include the generated power, ambient temperature, module temperature, and irradiation.

There will be two datasets used, one from temperature sensor data from two solar plants in India and the other from photovoltaic energy generation. 34 days' worth of generation data from 2020 are included in these databases. The analysis will be facilitated and made simpler by using data from just one of the photovoltaic plants.

**Dataset names:** Plant\_1\_Generation\_Data and Plant\_1\_Weather\_Data.

The project's objective is to discover how energy generation is related to environmental factors and then create a power generation prediction model using the collected data.

# Variable Identification

## Generation data

* DATE\_TIME - Date and time for each observation. Observations recorded at 15 minute intervals.
* PLANT\_ID - Plant ID number.
* INVERTER - Inverter id.
* DC\_POWER - Amount of DC power generated by the Inverter in this 15 minute interval (kW).
* AC\_POWER - Amount of AC power generated by the Inverter in this 15 minute interval (kW).
* DAILY\_YIELD - Daily yield is a cumulative sum of power generated on that day, till that point in time.
* TOTAL\_YIELD - This is the total yield for the inverter till that point in time.

## Temperature and Solar Irradiation data

* DATE\_TIME - Date and time for each observation. Observations recorded at 15 minute intervals.
* Plant ID - this will be common for the entire file.
* SENSOR - Stands for the sensor panel id.
* AMBIENT\_TEMPERATURE - This is the ambient temperature at the plant.
* MODULE\_TEMPERATURE - There is a module (solar panel) attached to the sensor panel. This is the temperature reading for that module.
* IRRADIATION - Amount of irradiation for the 15 minute interval.

# The Employed Technique

The method used in this project involves several steps:

1. **Data Selection:** Energy generation and weather data from two solar plants were selected from Kaggle, but only data from one of the plants was utilized. This decision was made to simplify the model and streamline the code.
2. **Exploratory Data Analysis (EDA):** EDA is employed to gain an understanding of the data structure, identify missing values, examine descriptive statistics, and visualize the relationships between variables.
3. **Data Preprocessing:** For further analysis, adjustments were made to variable types, handling of missing data, column renaming, and data aggregation into specific time intervals.
4. **Implementation of Linear Regression:** A linear regression model was implemented using the prepared data to predict energy generation using temperature and irradiation data.

# Exploratory Data Analysis

Energy generation and temperature data were thoroughly analyzed. Variable identification, missing data verification, descriptive statistical analysis, and visualization of variable relationships were all components of this process. Heatmaps, histograms, and scatter matrices were used to determine the relationships between variables.

## Descriptive Statistics - Generation data

DATE\_TIME DC\_POWER AC\_POWER \  
count 68778 68778.000000 68778.000000   
mean 2020-06-01 08:02:49.458256896 3147.426211 307.802752   
min 2020-05-15 00:00:00 0.000000 0.000000   
25% 2020-05-24 00:45:00 0.000000 0.000000   
50% 2020-06-01 14:30:00 429.000000 41.493750   
75% 2020-06-09 20:00:00 6366.964286 623.618750   
max 2020-06-17 23:45:00 14471.125000 1410.950000   
std NaN 4036.457169 394.396439   
  
 DAILY\_YIELD TOTAL\_YIELD   
count 68778.000000 6.877800e+04   
mean 3295.968737 6.978712e+06   
min 0.000000 6.183645e+06   
25% 0.000000 6.512003e+06   
50% 2658.714286 7.146685e+06   
75% 6274.000000 7.268706e+06   
max 9163.000000 7.846821e+06   
std 3145.178309 4.162720e+05

## Descriptive Statistics - Temperature and Solar Irradiation data

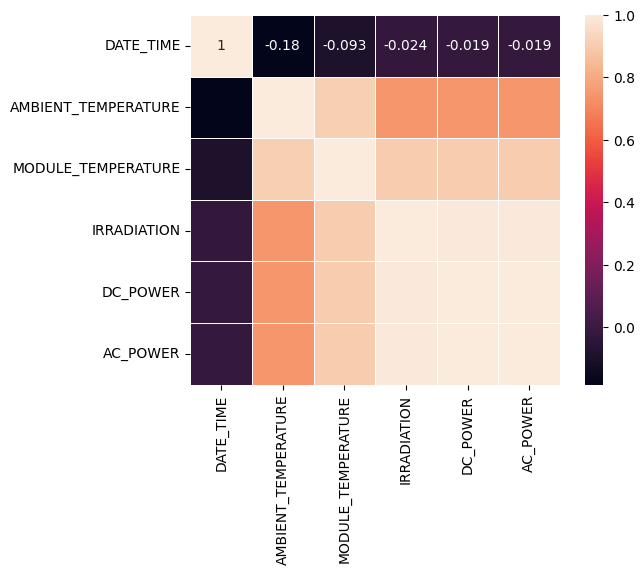
DATE\_TIME AMBIENT\_TEMPERATURE MODULE\_TEMPERATURE \  
count 3182 3182.000000 3182.000000   
mean 2020-06-01 05:52:22.080452608 25.531606 31.091015   
min 2020-05-15 00:00:00 20.398505 18.140415   
25% 2020-05-23 22:48:45 22.705182 21.090553   
50% 2020-06-01 09:52:30 24.613814 24.618060   
75% 2020-06-09 16:56:15 27.920532 41.307840   
max 2020-06-17 23:45:00 35.252486 65.545714   
std NaN 3.354856 12.261222   
  
 IRRADIATION   
count 3182.000000   
mean 0.228313   
min 0.000000   
25% 0.000000   
50% 0.024653   
75% 0.449588   
max 1.221652   
std 0.300836

## Merging Dataframes

Merging the plant1\_sensor and plant1\_data DataFrames using the 'DATE\_TIME' column as the key. This results in a new DataFrame called power\_sensor that contains all columns from both DataFrames, but combined based on the timestamp. Removing redundant columns from the power\_sensor DataFrame, which were created as a result of merging the DataFrames. These columns are removed because they are duplicates.

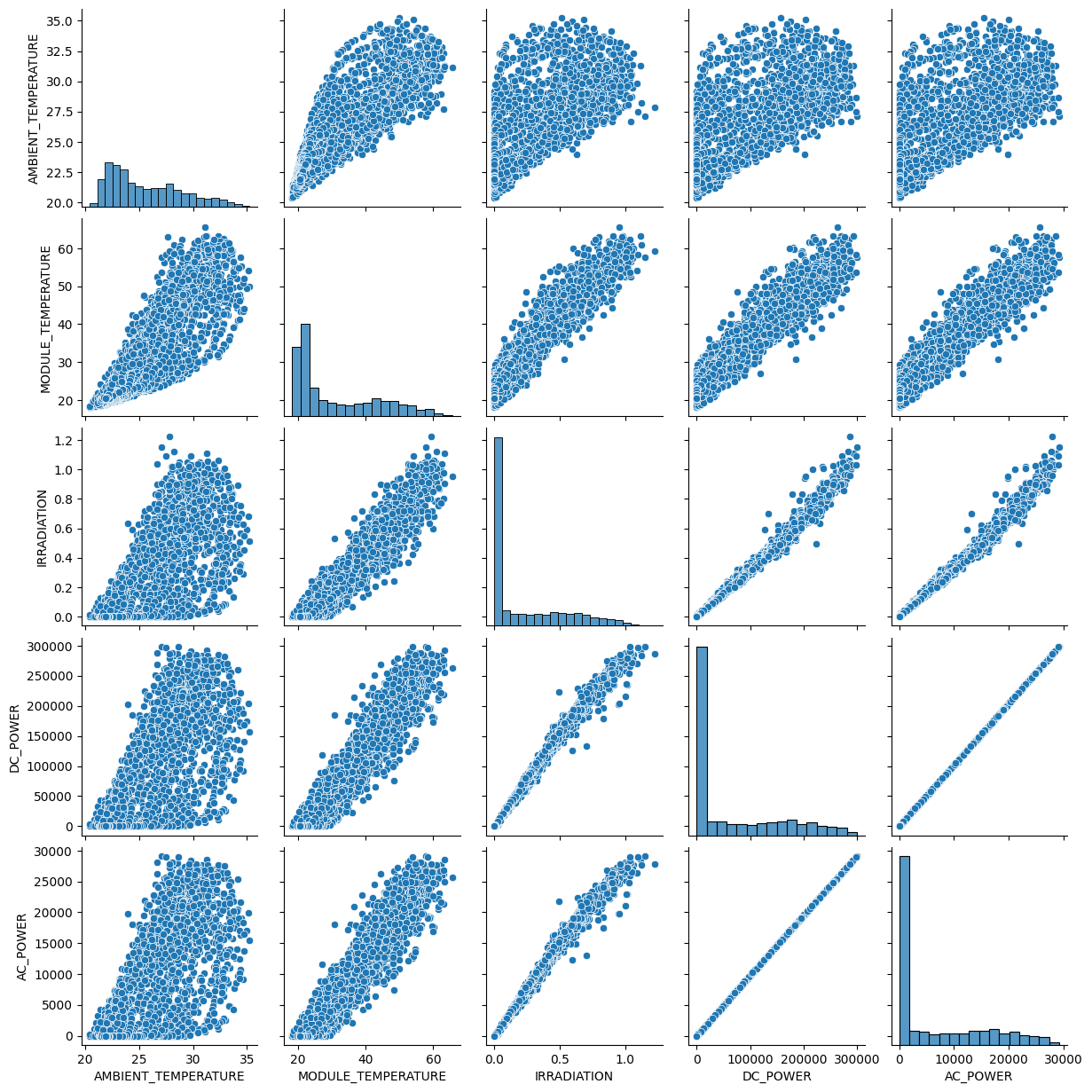
## Heatmap

By merging the two tables, it is possible to discover correlations between the different variables.



### Matrix of Scatter Plots

By merging the two tables, It allows identifying linear or non-linear relationships between pairs of variables and helps in detecting patterns, trends, and outliers in the solar plant's generation and temperature data.

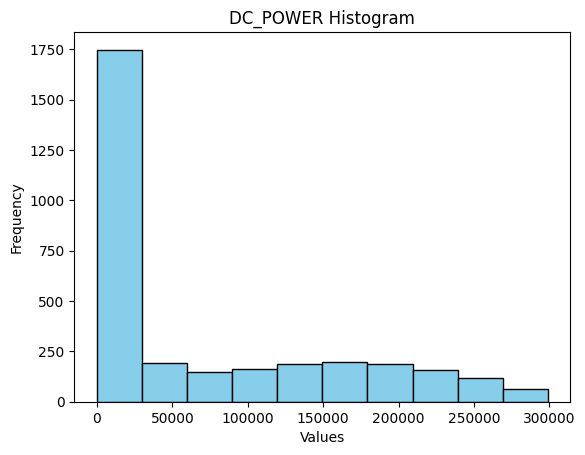


# Data Processing Forecasting

To facilitate analysis, preprocessing steps included data type conversion, handling of missing data, and aggregation of data into longer time intervals.

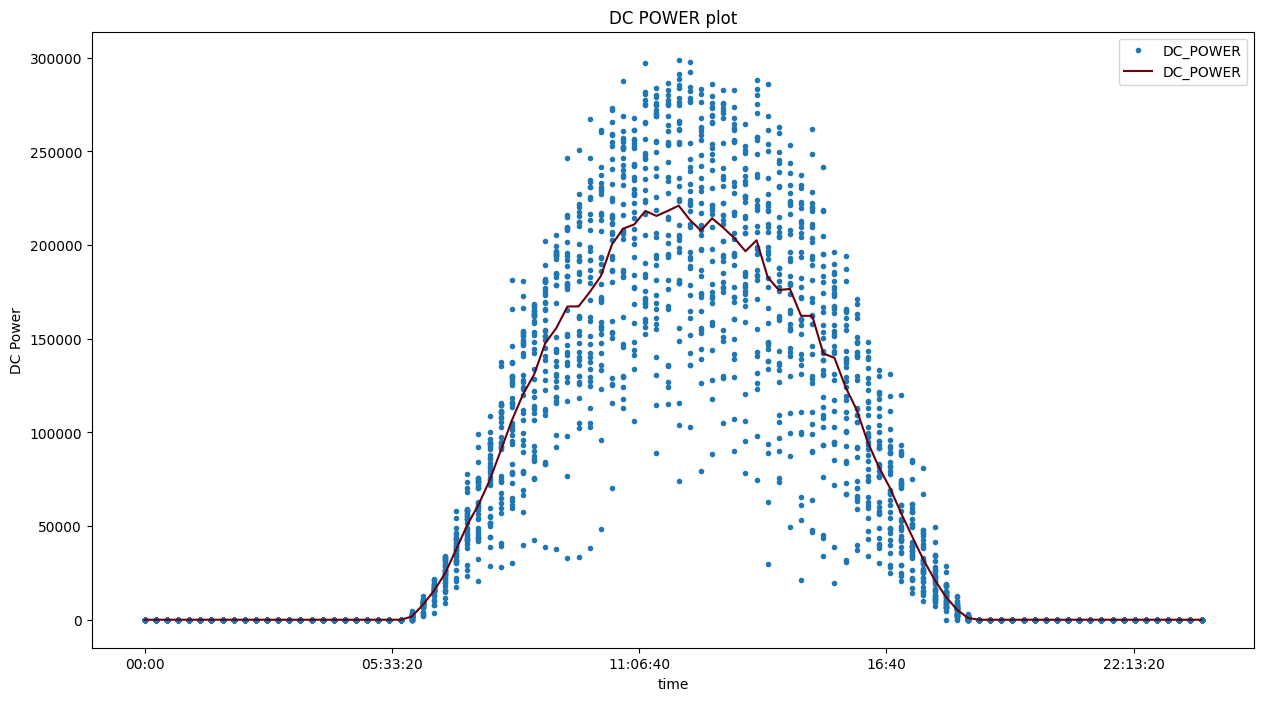
## DC Power

### Histogram



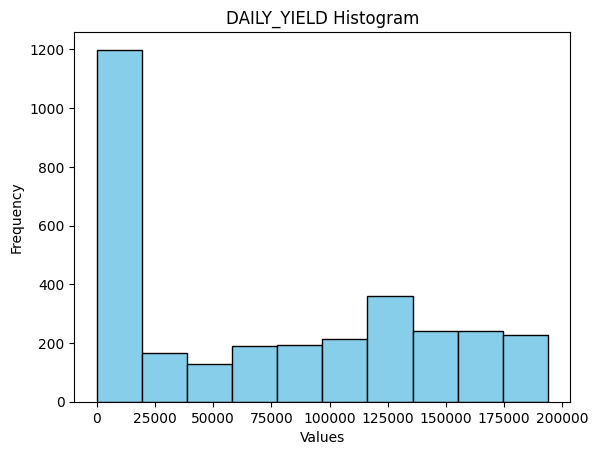
### Scatter Plot and Mean DC Power

Plot the data points from the 'DC\_POWER' column against the 'time' column in the format of a scatter plot and plot the mean of DC power on a line graph.



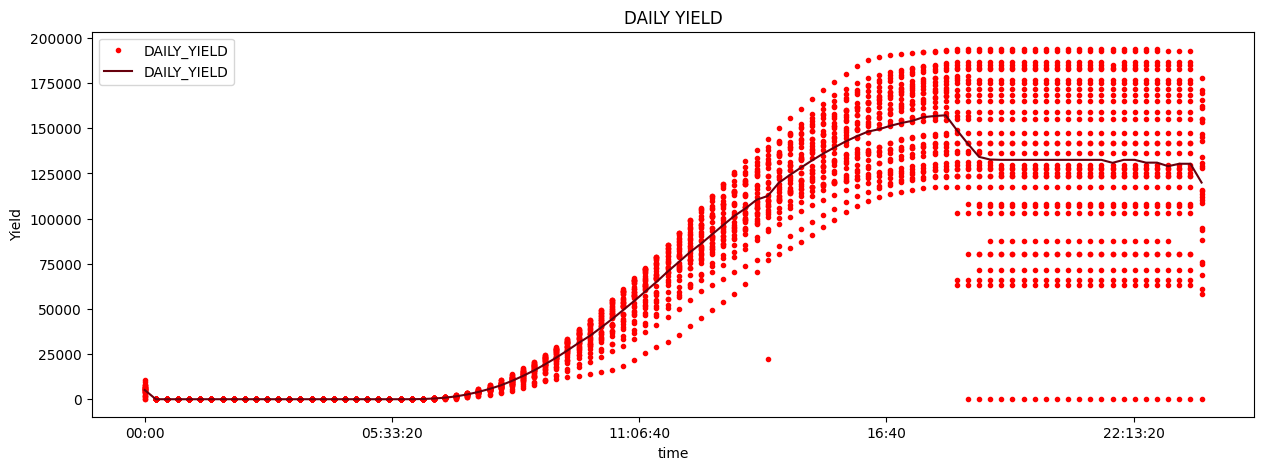
## Daily Yield

### Histogram



## Scatter Plot and Yield Mean

Plot the data points from the 'DAILY\_YIELD' column against the 'time' column in the format of a scatter plot and plot the mean of daily yield on a line graph.



## Daily Yield and AC-DC Power

Transform the original data into a new DataFrame (df\_daily\_gen), grouping the values for each time interval and adding an additional column ('time') to allow time-based analysis.

## 

## Module Temperature

## 

## Scatter Plot and Mean Temperature

# 

# Application of Linear Regression

Using temperature and irradiation data, a linear regression model was used to predict energy generation. The coefficient of determination (R2), mean squared error (MSE), and mean absolute error (MAE) were measured to evaluate the model's performance.

## Data Information

power\_sensor is the DataFrame containing the merged data of power generation and temperature from the Solar Plant. These data will be used in the machine learning model.

<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 3157 entries, 0 to 3156  
Data columns (total 8 columns):  
 # Column Non-Null Count Dtype   
--- ------ -------------- -----   
 0 DATE\_TIME 3157 non-null datetime64[ns]  
 1 AMBIENT\_TEMPERATURE 3157 non-null float64   
 2 MODULE\_TEMPERATURE 3157 non-null float64   
 3 IRRADIATION 3157 non-null float64   
 4 DC\_POWER 3157 non-null float64   
 5 AC\_POWER 3157 non-null float64   
 6 DAILY\_YIELD 3157 non-null float64   
 7 TOTAL\_YIELD 3157 non-null float64   
dtypes: datetime64[ns](1), float64(7)  
memory usage: 197.4 KB

## Training and Test Sets to Linear Regression Model

Selecting three columns from the power\_sensor DataFrame to be our input features (features) and selecting the 'DC\_POWER' column from the power\_sensor DataFrame to be our prediction target.

## Coefficient of Determination ( R^2 )

Calculates the coefficient of determination ( R^2 ) (R-squared) to evaluate the performance of the linear regression model. It takes as input the actual values from the test set and the predictions made by the model. The closer to 1, the better the model is at explaining the variability of the data.inear.

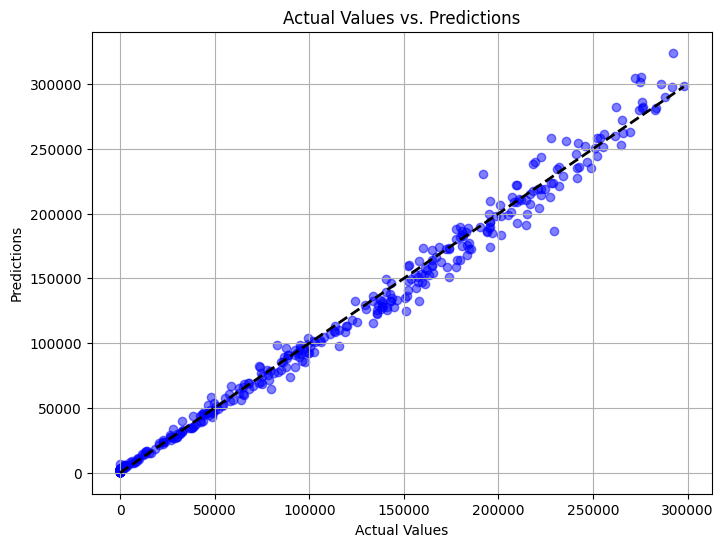
**Coefficient of Determination (R²): 0.9939010484064352**

## MSE and MAE

* MSE is a measure of the average squared error between predictions and actual values. The lower the MSE, the better the model is at fitting the data.
* MAE is a measure of the average absolute differences between predictions and actual values. The lower the MAE, the better the model's performance.

**Mean Squared Error (MSE): 45216830.68690255**

**Mean Absolute Error (MAE): 4000.6956272785133**

****

# Findings

The coefficient of determination (R2) was 0.9939 (99,39%), the mean squared error (MSE) was 45216830.6869, and the mean absolute error (MAE) was 4000.6956. This indicates that the model is extremely accurate in predicting energy generation using temperature and irradiation data.

# Conclusions and Suggestions for the Future

The project's goal of utilizing a linear regression model to forecast energy generation by analyzing temperature and photovoltaic energy generating data from an Indian solar plant was accomplished.

Data on temperature and energy generation are highly correlated. Future proposals include investigating different machine learning models, including neural networks, to further enhance predictions. Additionally, gathering more information on humidity and geographic location might help prediction models become more accurate.

In order to find equipment with low efficiency, we can also examine other interactions in the future, such as the one between generation and the inverter that generated it.

# References

Iqbal, M., Linear Regression - Week 4. Study Material, CCT College Dublin, Dublin.

McQuaid, D. What is Exploratory Data Analysis. Study Material, CCT College Dublin, Dublin.

# GitHub Link

<https://github.com/FelipeSteudel/Felipe_Steudel-Strategic_Thinking-CA2.git>