1. Introduction

(1) This report aims to assist the government in creating a tool to help users predict solar panel power generation. The specific approach is to use Python to establish a model, and when users input information about their houses, the model can generate estimated predicted electricity generation for users' reference.

(2) To assist the government in creating a tool to help users predict solar panel power generation. The specific approach is to use Python to establish a model based on government collected data. When users input information about their houses, the model can generate estimated predicted electricity generation. Predicting power generation can provide users with data support to determine whether solar panels are suitable for installation.

(3) The database is survey data on solar panels, including 3000 sets of data and 13 variables. The data comes from energy retailers, energy distributors, and solar installers. It contains various types of data, such as Household ID, City name, Latitude, House Type, Roof Type, and so on.

(4) Three models were established for predicting power generation, and the M3 model with the smallest MSE was selected as the optimal model based on the OLS method. The M3 model was compared with Benchmark Model 1 and Benchmark Model 2, and ultimately M3 was selected as the optimal model for power generation prediction.

2. Candidate model

Model 1: x:(Panel_Capacity, Latitude)

$y = \beta 0 + \beta 1 * Panel\_Capacity + \beta 2 * Latitude + \epsilon i$

Model 2: x:(Panel_Capacity, Latitude, None, Significant)

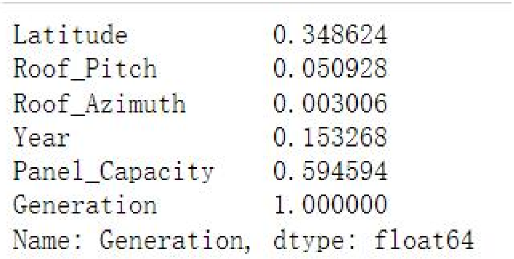
$y = \beta 0 + \beta 1 * Panel\_Capacity + \beta 2 * Latitude + \beta 3 * None + \beta 4 * Significant + \epsilon i$

Model 3: x:(Panel_Capacity, Latitude, None, Significant, Azimuth0-45, Azimuth45-90, Azimuth90-135)

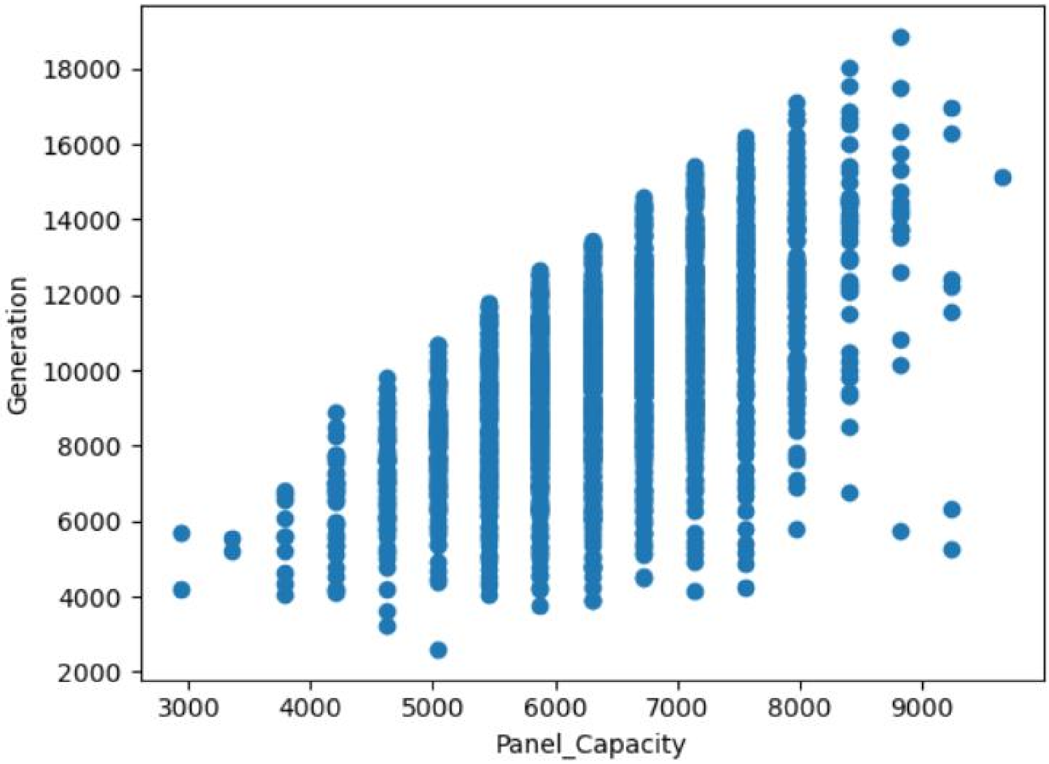
$y = \beta 0 + \beta 1 * Panel\_Capacity + \beta 2 * Latitude + \beta 3 * None + \beta 4 * Significant + \beta 5 * Azimuth0\text{-}45 + \beta 6 * Azimuth45\text{-}90 + \beta 7 * Azimuth90\text{-}135 + \epsilon i$

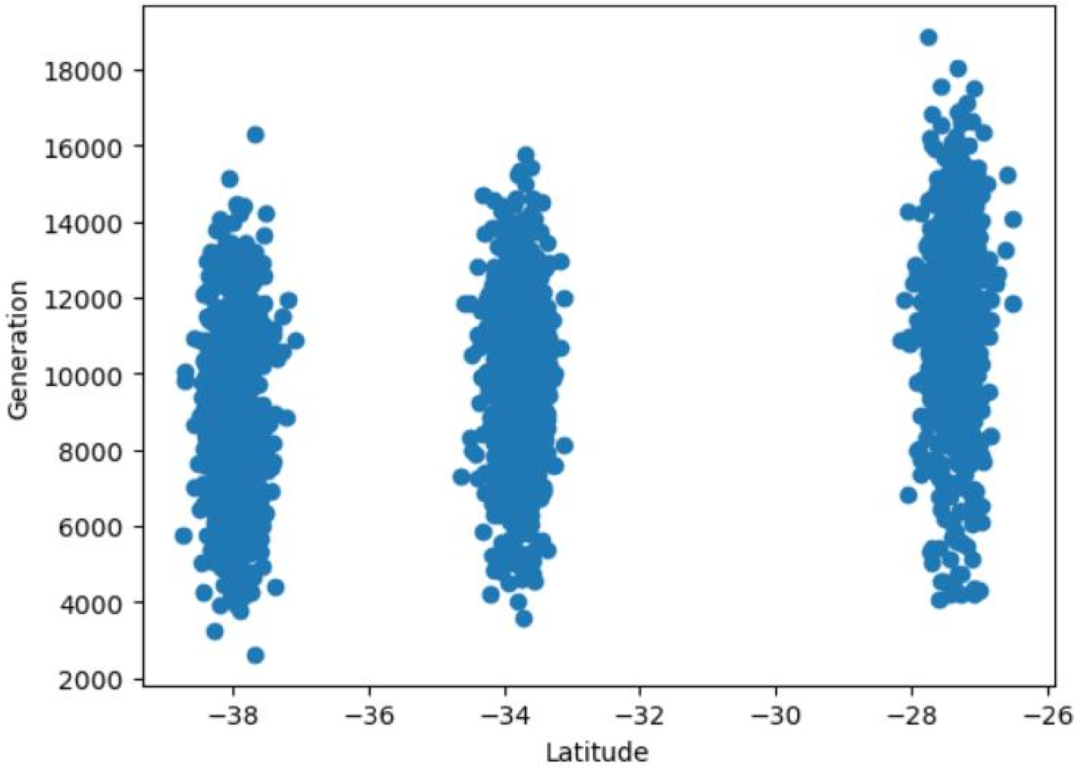
Reasons why choose xi:

Panel_Capacity & Latitude :



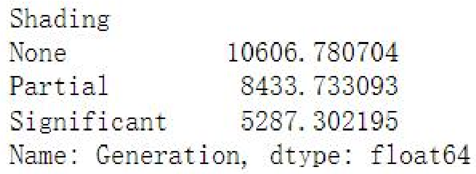
```
Latitude           0.348624
Roof_Pitch         0.050928
Roof_Azimuth       0.003006
Year               0.153268
Panel_Capacity     0.594594
Generation         1.000000
Name: Generation, dtype: float64
```



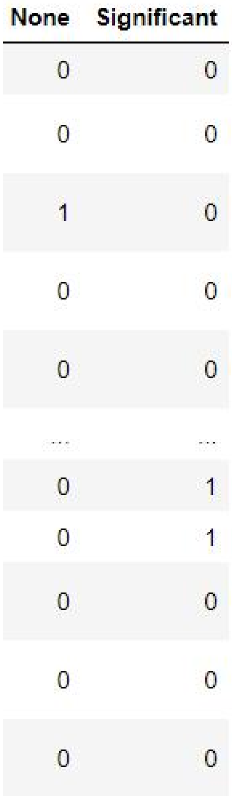


After data cleaning, we explored the correlation of generation and found that the two highest correlation values were Panel_ Capacity and Latitude, therefore they may be used to predict Generation.By conducting scatter plot analysis on these two independent variables and generation, it can be found that there is a certain correlation between the independent variables and generation.

None & Significant:



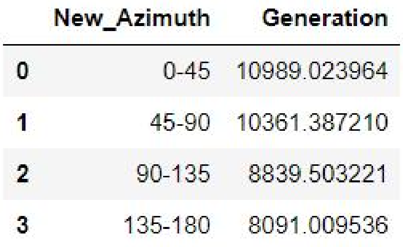
```
Shading
None           10606.780704
Partial         8433.733093
Significant     5287.302195
Name: Generation, dtype: float64
```



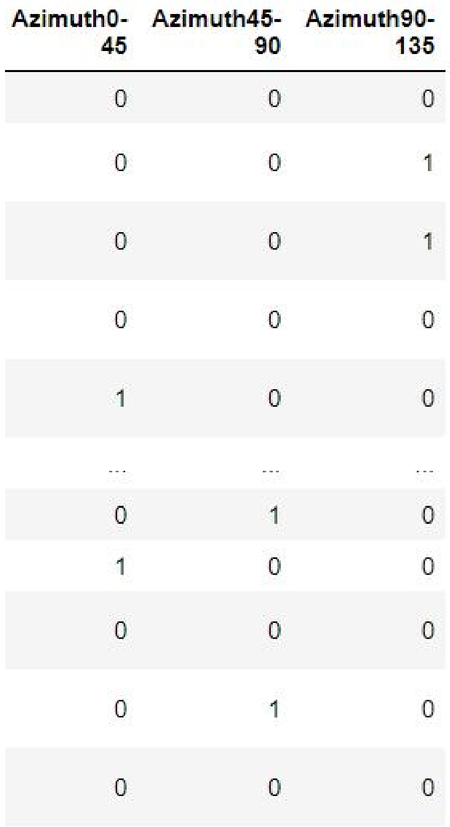
None	Significant
0	0
0	0
1	0
0	0
0	0
...	...
0	1
0	1
0	0
0	0
0	0

Compared to other categorical variables, Shading has a significant relationship with Generation. When Shading is None, the value of Generation is significantly higher. When Shading is Significant, the opposite is true. When Partial, the value of Generation is at a moderate level. Therefore, perform dummy variable processing on Shading.

Azimuth0-45, Azimuth45-90, Azimuth90-135:



	New_Azimuth	Generation
0	0-45	10989.023964
1	45-90	10361.387210
2	90-135	8839.503221
3	135-180	8091.009536



Azimuth0-45	Azimuth45-90	Azimuth90-135
0	0	0
0	0	1
0	0	1
0	0	0
1	0	0
...	...	...
0	1	0
1	0	0
0	0	0
0	1	0
0	0	0

These variables are related to the Roof_ The result of Azimuth's angle classification processing followed by dummy variable processing. Right Roof_ Azimuth was divided into four groups and found significant correlations between different perspectives on Generation. Therefore, perform dummy variable processing on the sub type data.

There are four dimensions of evaluation criteria for error, namely Linearity, Independence, Normality, and Equal Variance.

Assumptions: The relationship between all xi and generation is linear. The error is completely independent and not affected by other variables. The data distribution of errors belongs to the type of normal distribution. And the error has equal variance.

By using the Ordinarily Least Squares (OLS) method, the sum of squared residuals between the predicted values of the model and the actual observations is minimized to select the optimal regression coefficient β i.

3. Model estimation and selection

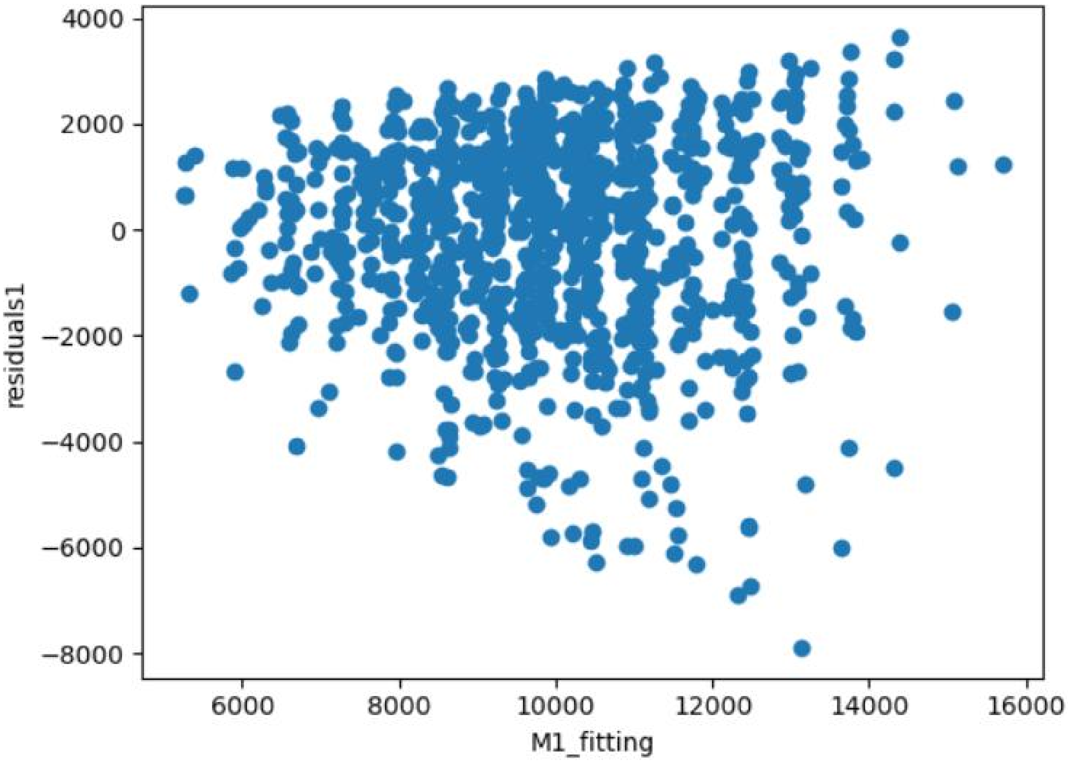
(1) Perform model estimation in the train stage, model selection in the valid stage, calculate mse, and find the optimal model with the smallest mse. Run the optimal model into the train and valid stages, and finally conduct model evaluation in the test stage.

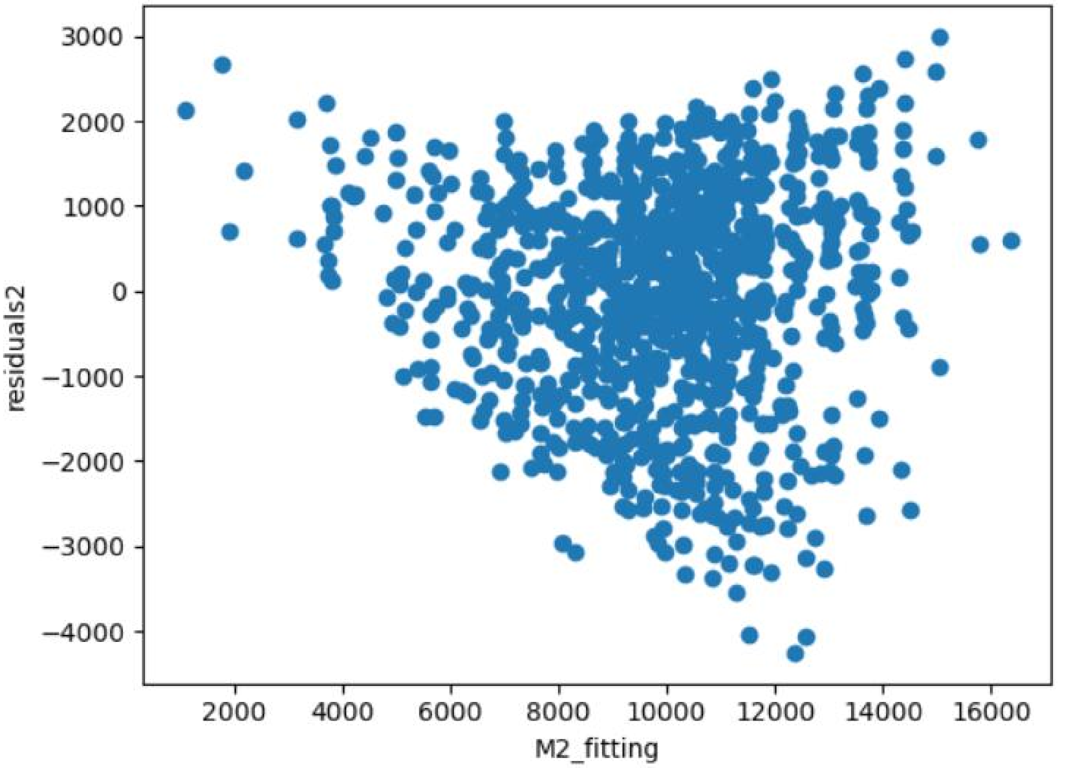
(2) Based on the training set, the parameters of model one are 1.55751622e+00, 2.34389246e+02, -2.23291587e+03, -5.47404868e+03, and the intercept is 7695.637225959193, respectively

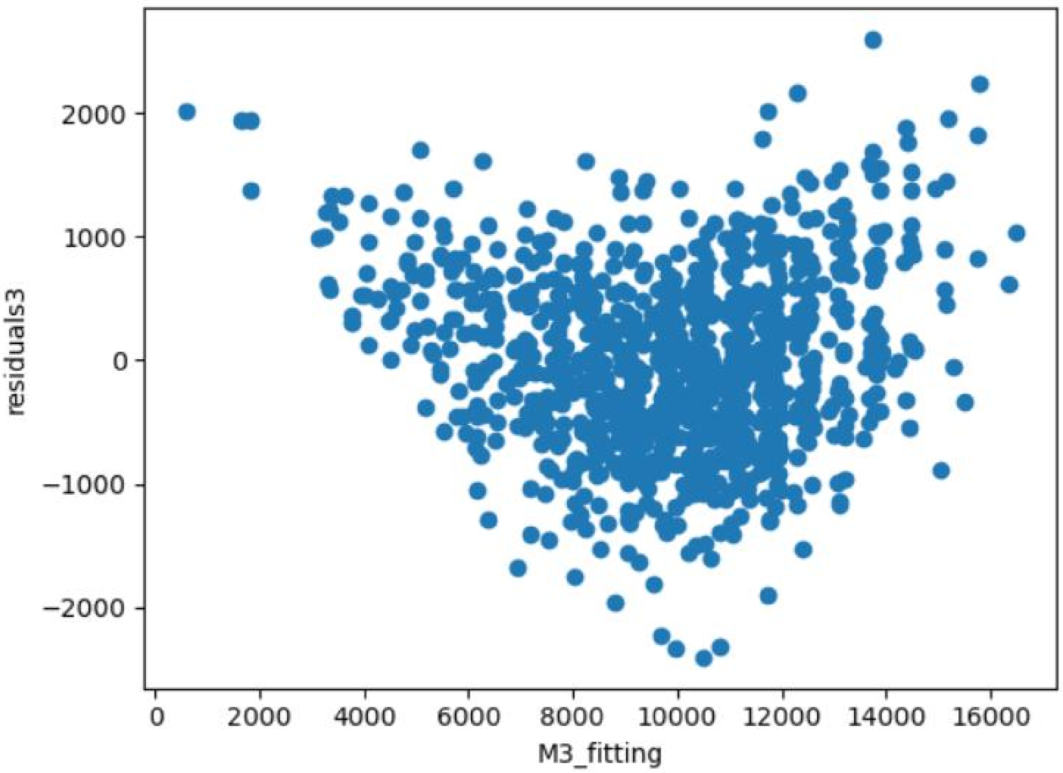
The parameters of Model 2 are 1.55751622e+00, 2.34389246e+02, -2.23291587e+03, -5.47404868e+03, and the intercept is 8350.47730107541. The parameters of Model 3 are

1.53284475e+00, 2.08333911e+02, -2.10930185e+03, -5.10797537e+03, respectively, -7.58579993e+02, -1.89701621e+03, -2.78181421e+03, intercept 8605.521642414642

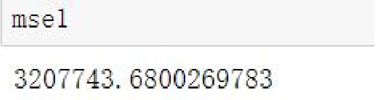
(3)







The above three figures are scatter plots of residual data for M1, M2, and M3, from the scatter plot, it can be seen that the variances are not equal. The other points are consistent with the assumptions.

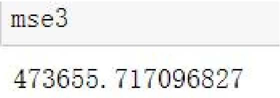


```
mse1

3207743.6800269783
```

```
mse2

1465290.576585396
```



```
mse3

473655.717096827
```

(4) The MSE of the first model is 3207743.6800269783, the MSE of the second model is 1465290.576585396, and the MSE of the third model is 473655.717096827.

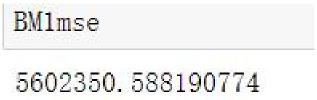
(5)The third model has the smallest MSE, therefore the third model is the best mode.

(6)The first model has 2 variables, the second model has 4 variables, and the third model has 7 variables Their MSE gradually decreases and their variance gradually increases, bias gradually decreases, so the first model is the simple, the third model is the most complex, and the second model is centered.

4. Model evaluation

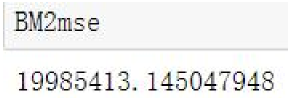
This report selects the third model as the optimal model, and the other optimal model is fitted in the steps of training and validation. Operate Benchmark Model 1 and Benchmark Model 2 using the same method to obtain the MSE of the three models, and select the model with the smallest MSE as the new optimal model.

Benchmark Model 1 contains generation data for different cities, predicting power generation based on different cities. Benchmark Model 2 contains the generation data corresponding to City and Panel Capacity, which is an additional variable compared to Benchmark Model 1 for predicting power generation.



```
BM1mse
```

```
5602350.588190774
```



```
BM2mse
```

```
19985413.145047948
```

After testing, it was found that the MSE of M3 is 473655.717096827, Benchmark Model 1 is 5602350.588190774, and Benchmark Model 2 is 19985413.145047948. Therefore, for the prediction of Generation, M3 has the best performance, followed by Benchmark Model 2, and finally Benchmark Model 1.

5. Conclusion

The finding of this report includes the early EDA process, which uses correlation to find the most suitable independent variable for predicting power generation. It also includes model testing and selection, and selecting the M3 model with the smallest error as the prediction model is the best model. The limitation of this report is that it only considers linear regression models and does not consider the selection and comparison of other models. In future work, we will consider expanding the number and types of models, comparing them, and finding the most suitable model. Meanwhile, through extensive training, learn the variables and data that are most suitable for predicting the model.