**Wind Speed Prediction Report**

**October 2018**

1. **Introduction**

The goal of this project is to predict the actual wind speed of windmills.

1.1 Data Introduction

There are 113866 rows rare data in train data, and over 8000 rows data in test data. There are 3 weather metrics including more than 15-dimension data. For each whether metric, they include wind speed, wind direction, temperature, wind density and atmospheric pressure. The target of attribute is Y.ws\_tb that means the wind speed on windmills. We firstly look at data condition of the target attribute as following table 1.

Table 1. the brief of speed of wind (Y.ws\_tb)

|  |  |
| --- | --- |
| count | 58886 |
| mean | 7.594716 |
| std | 2.807288 |
| min | 0.00825 |
| 25% | 5.48482 |
| 50% | 7.008785 |
| 75% | 9.19358 |
| max | 23.2954 |

From the above table, we could see that there is count 58886, and hence many empty values on Y.ws\_tb. Besides, the minimum value is about 0.008, but maximum value is over 23.

1.2 Data Exploration

In this section, we will explore data from as many aspects as possible. It mainly includes normal distribution of data, correlation among data, scatter plot, box plot and honeycomb graph of kinds of data.

1.2.1 Inspection of Normal Distribution

Draw the distribution graph of the Y.ws\_tb as Figure 1. Here, blue line is data distribution and black line is fitting normal distribution. Sometimes, the data following normal distribution could help well to make the prediction. Hence, consider transform data at pre-processing stage. And the graph of distribution after processing as Figure 2.

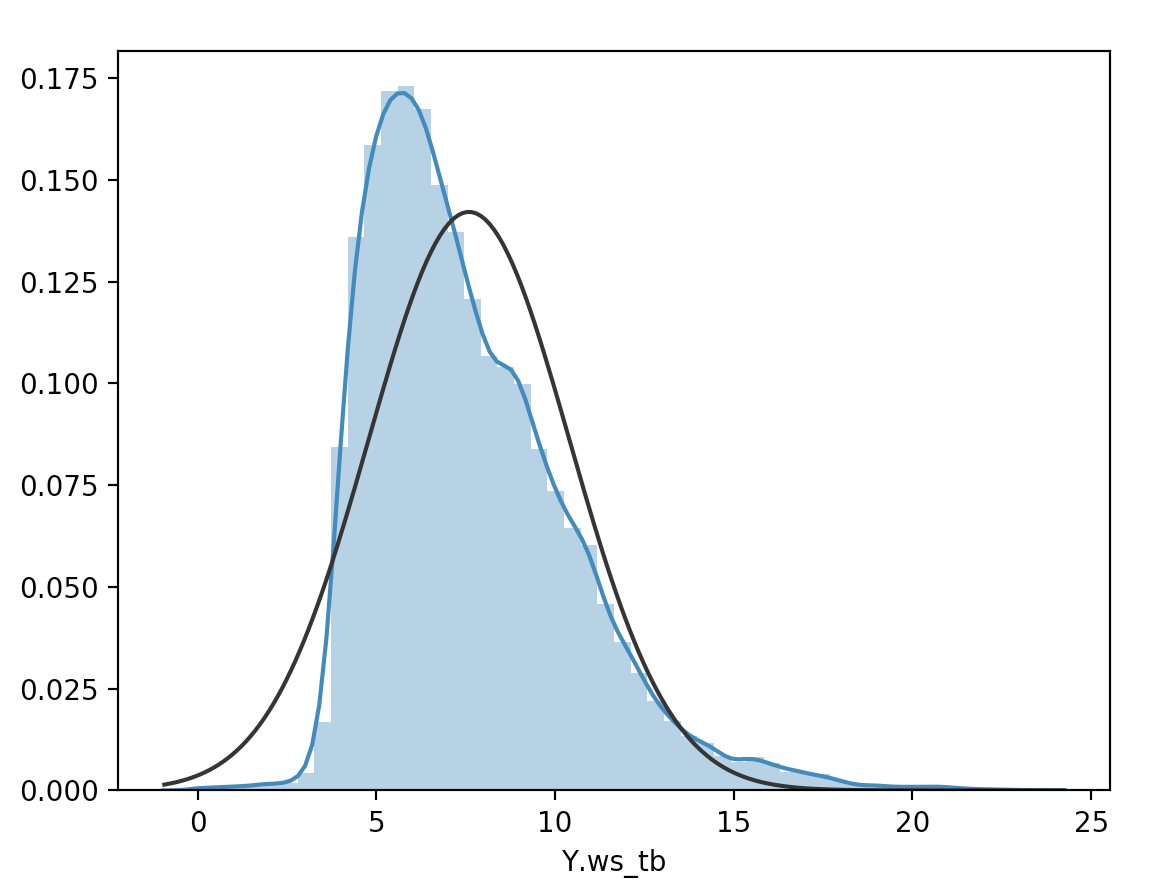


Figure 1. Distribution of Y.ws\_tb (Before)

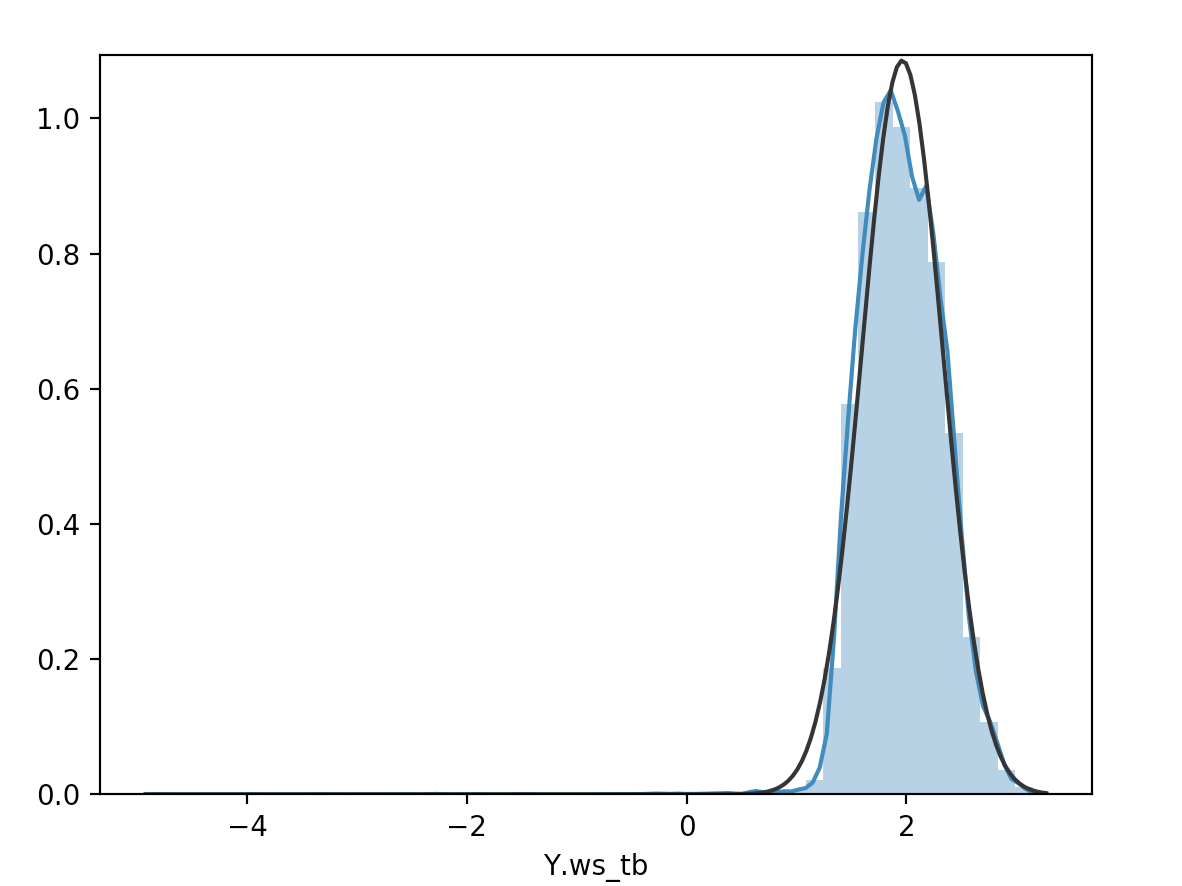
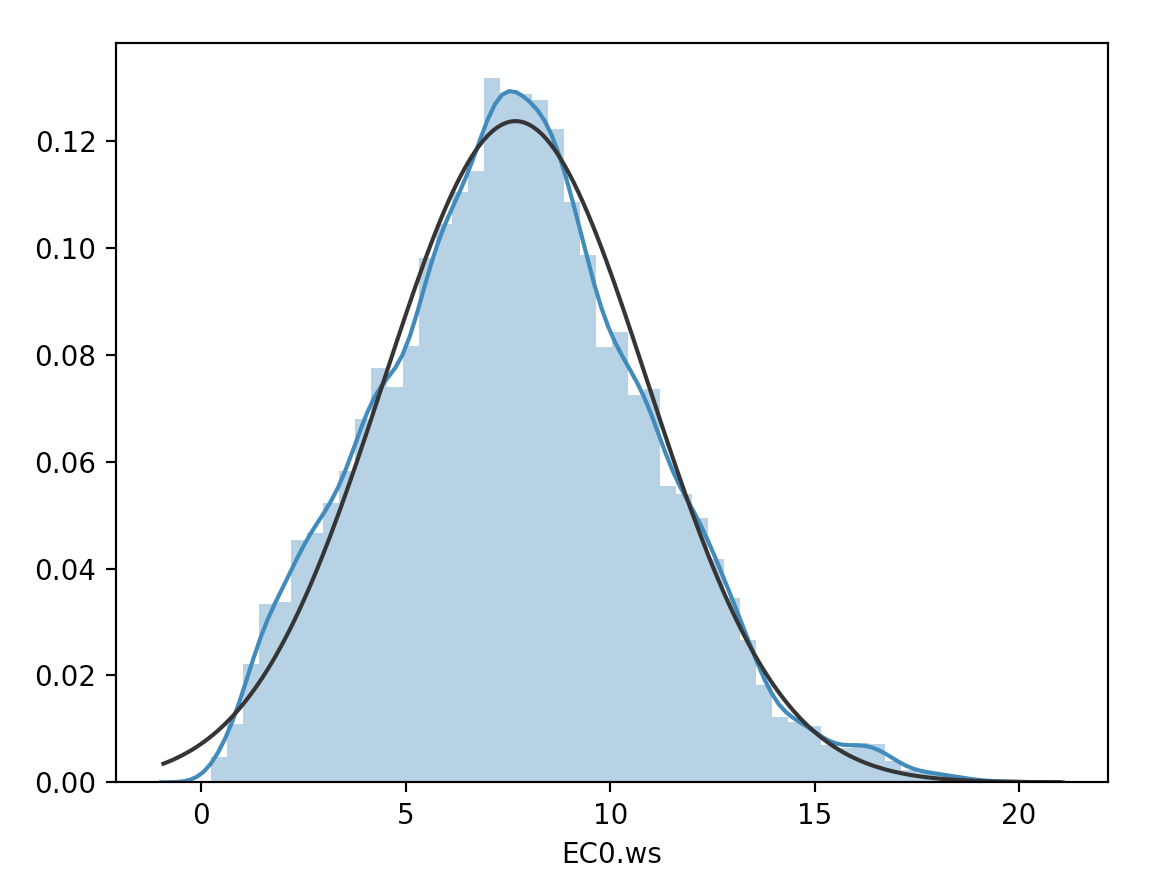
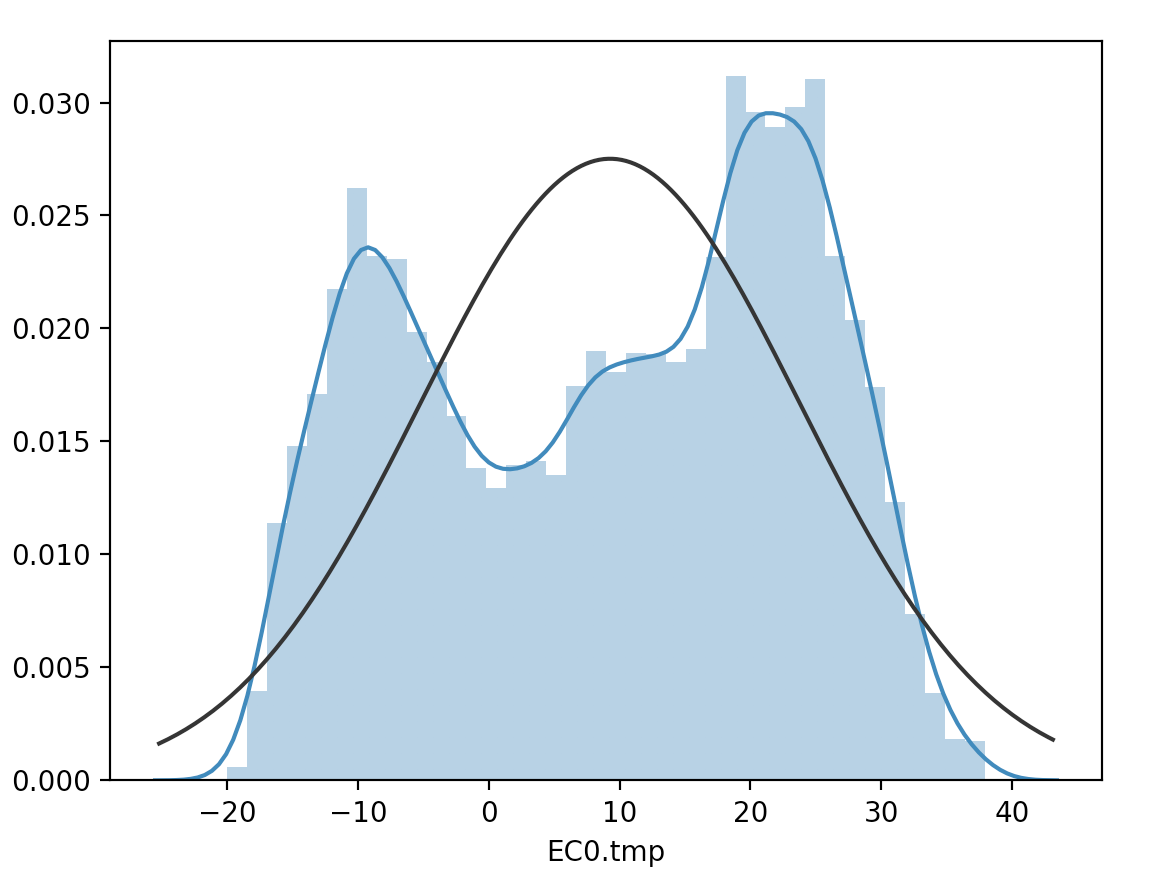
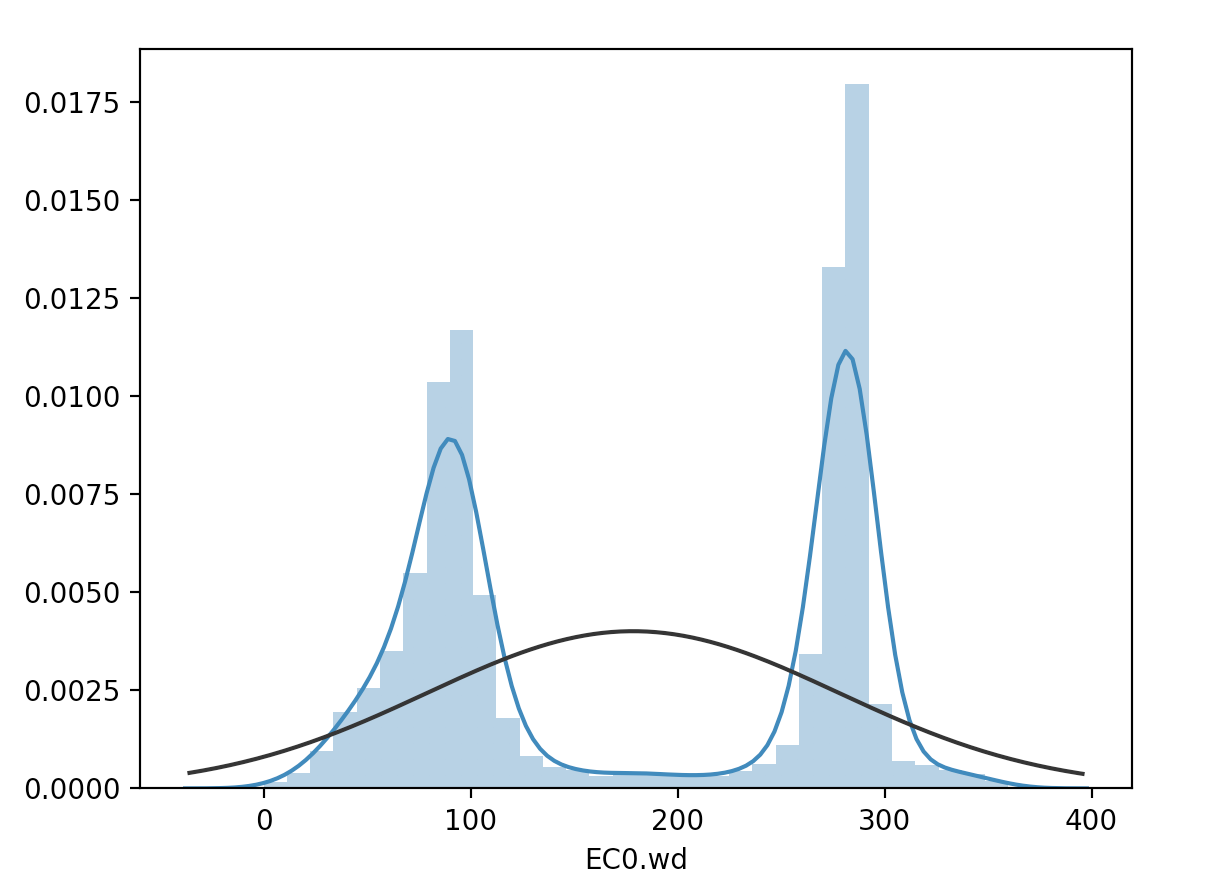


Figure 2. Distribution of Y.ws\_tb (After)

Besides the target attribute, the distribution of other important attributes from three weather sources as followings.





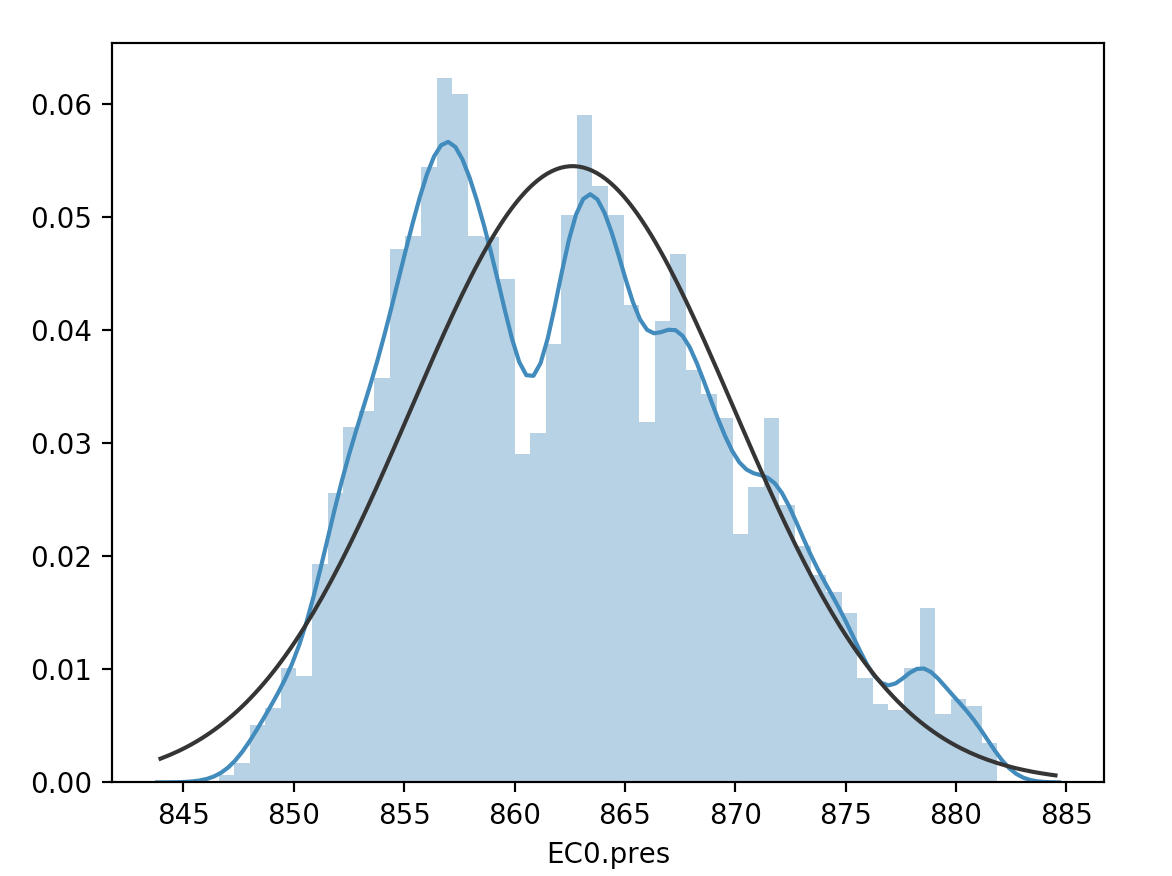
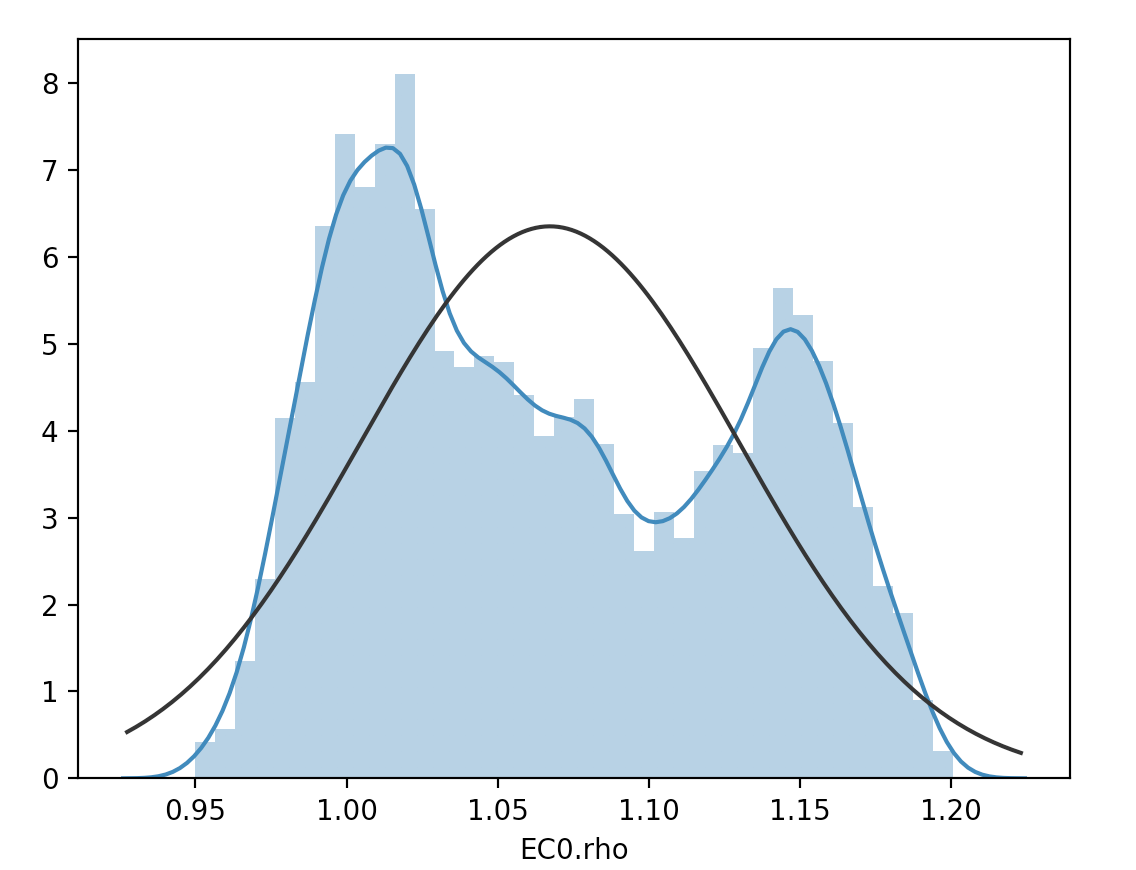
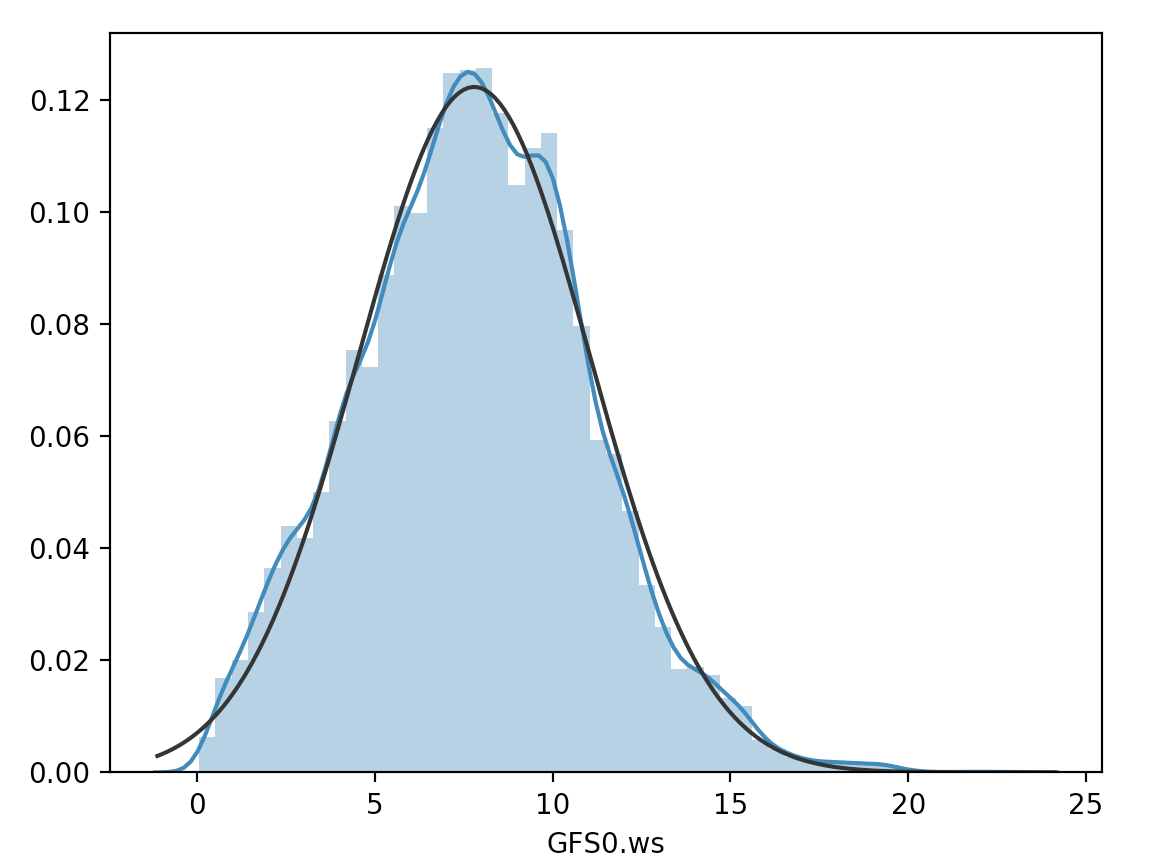
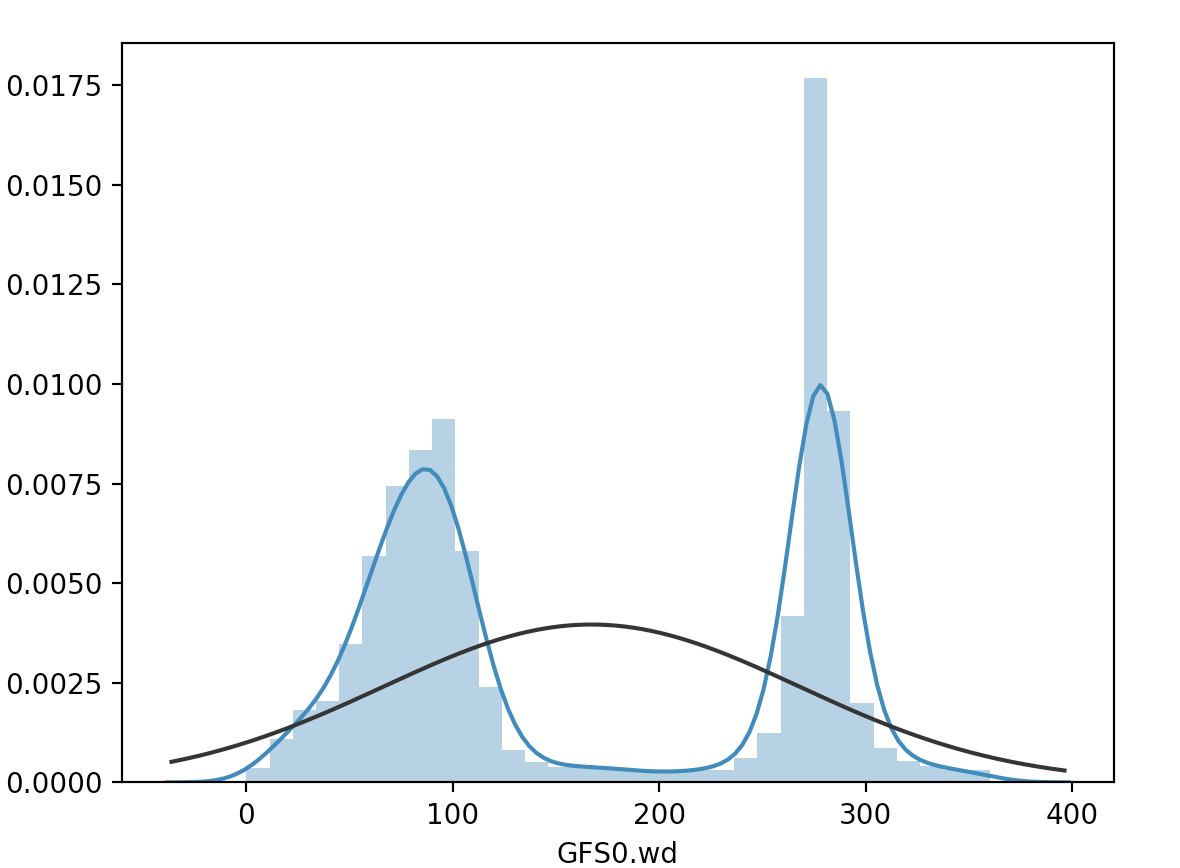


Figure 3. The Distribution of EC0 Attributes





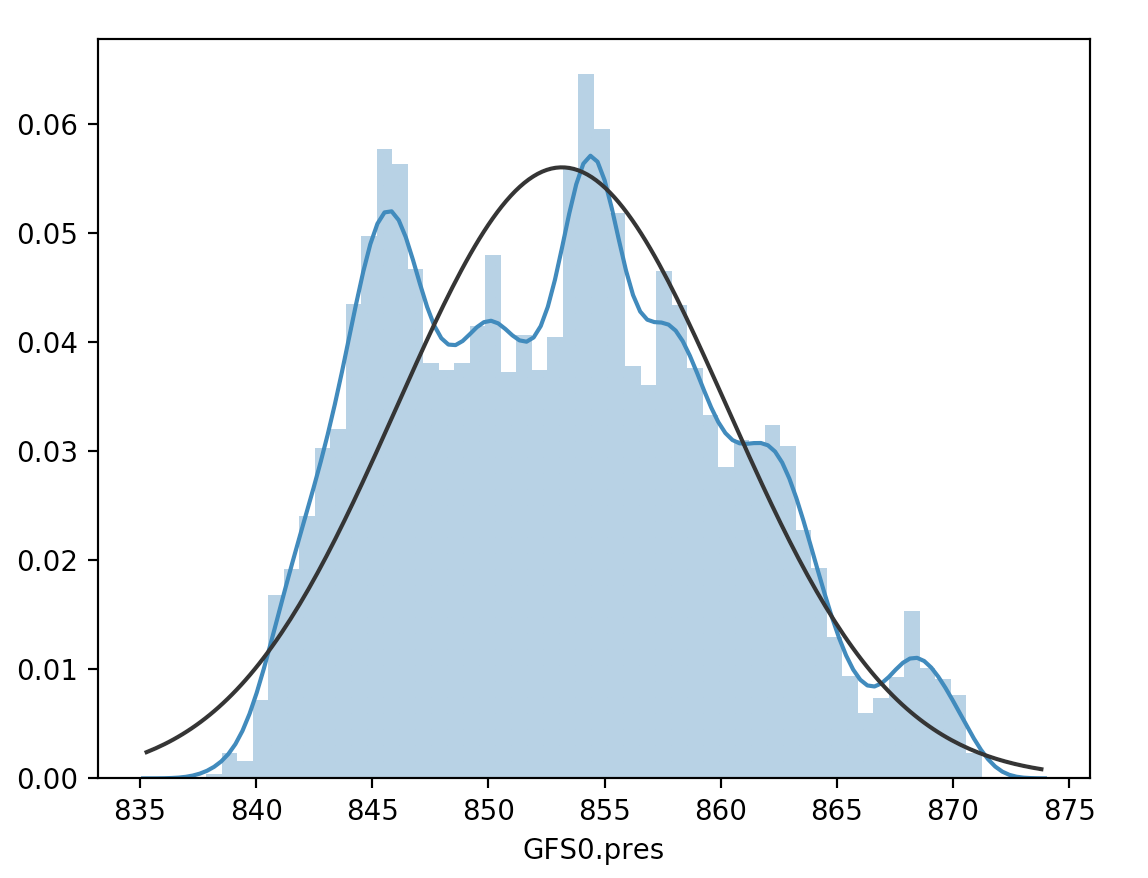
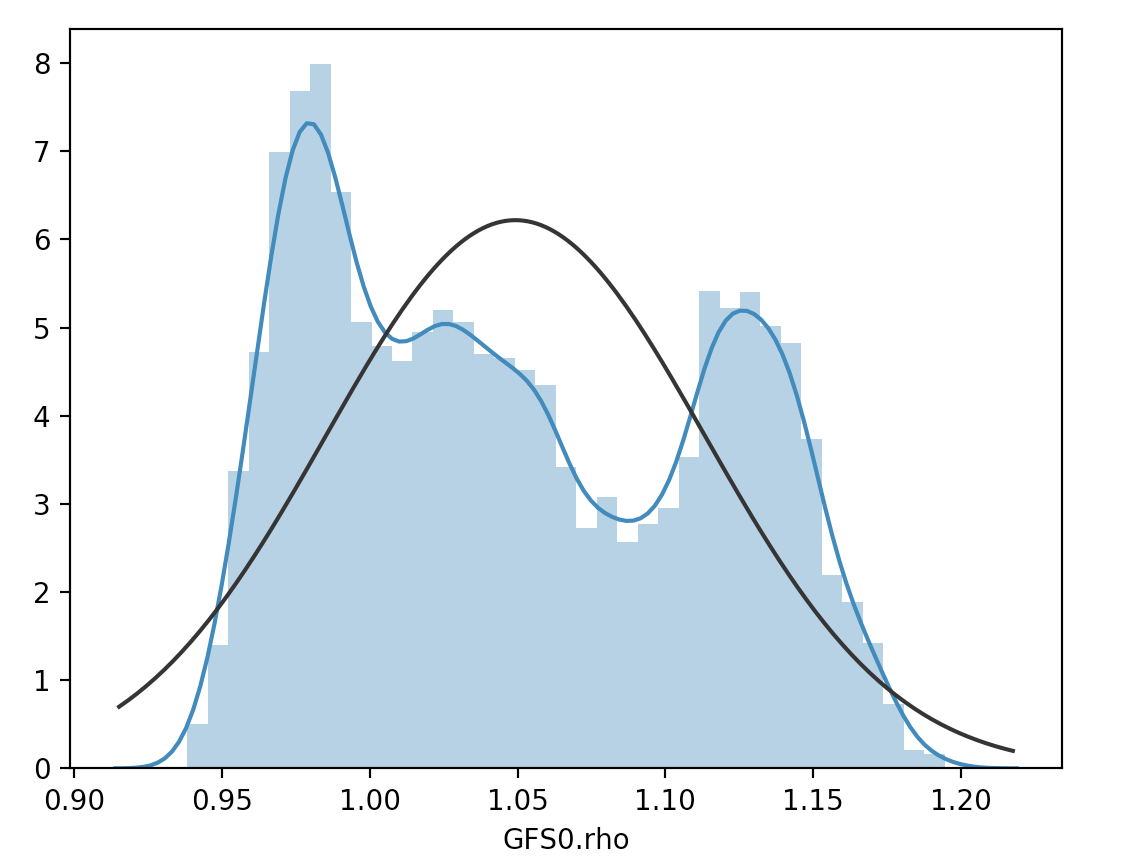
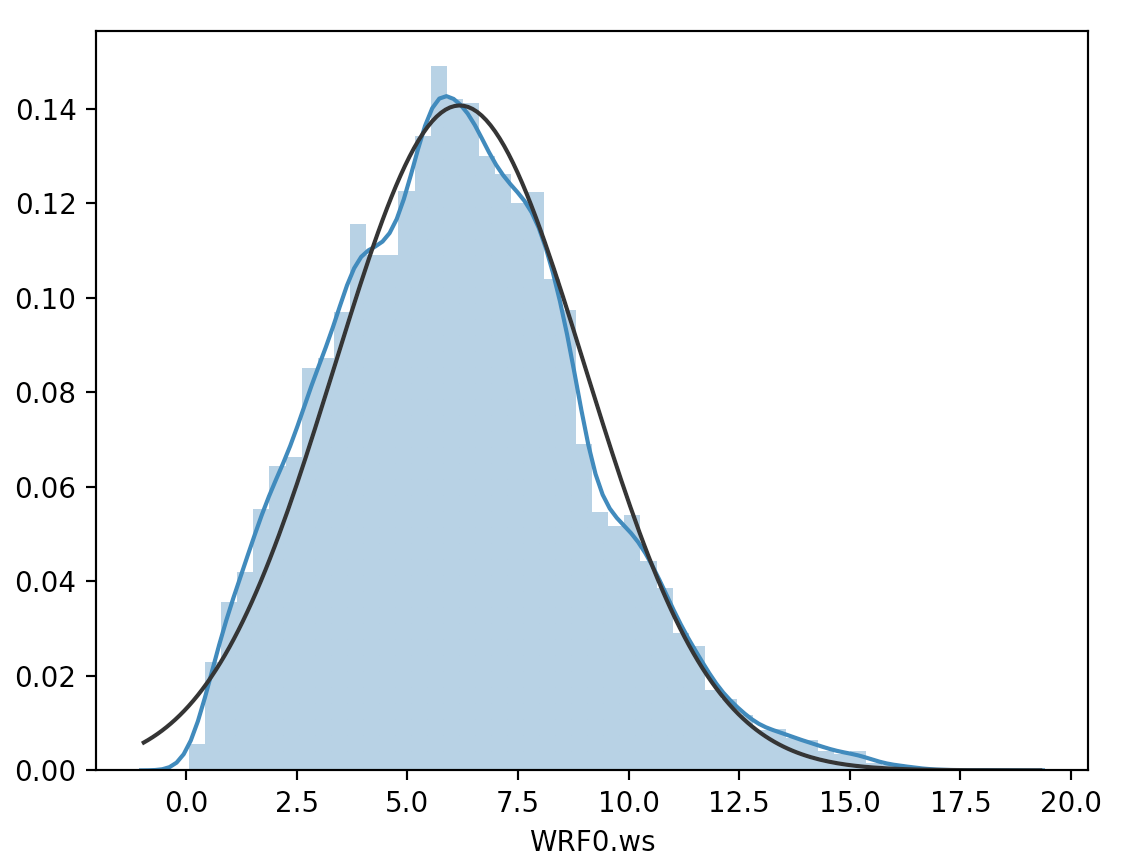
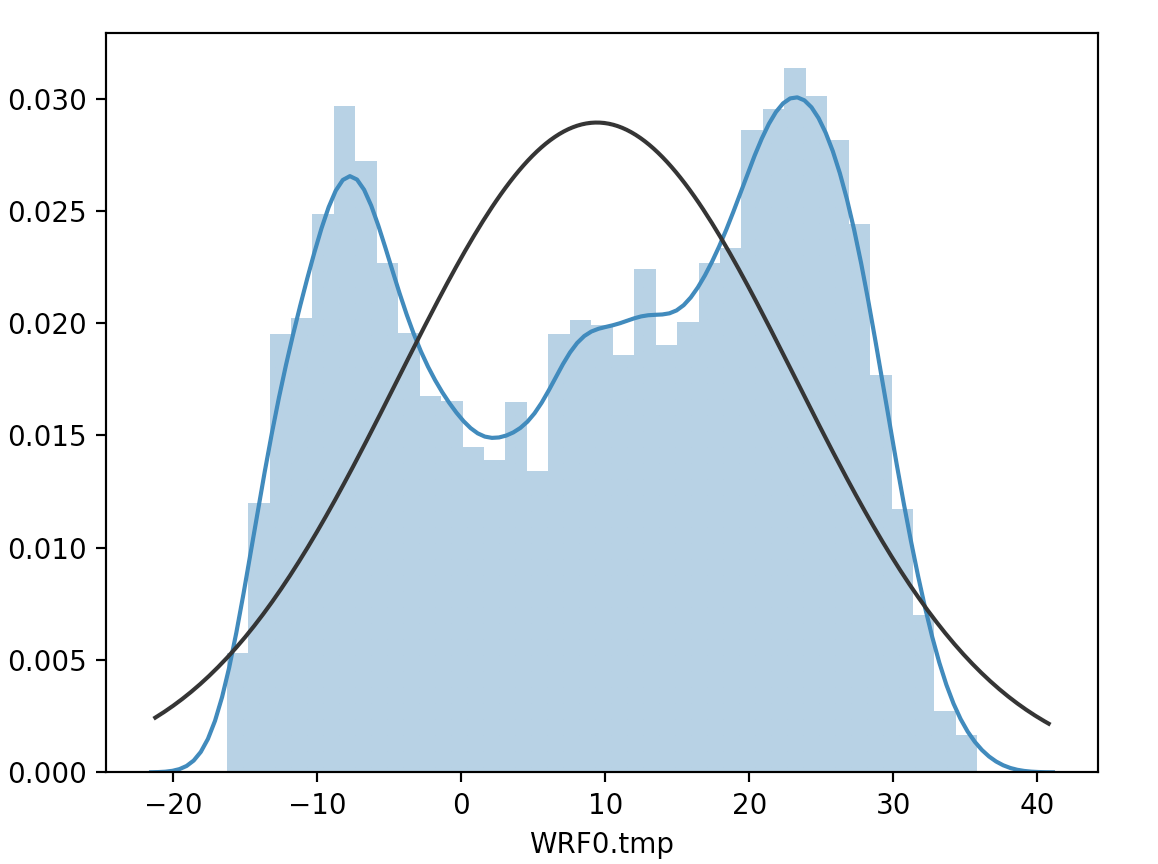
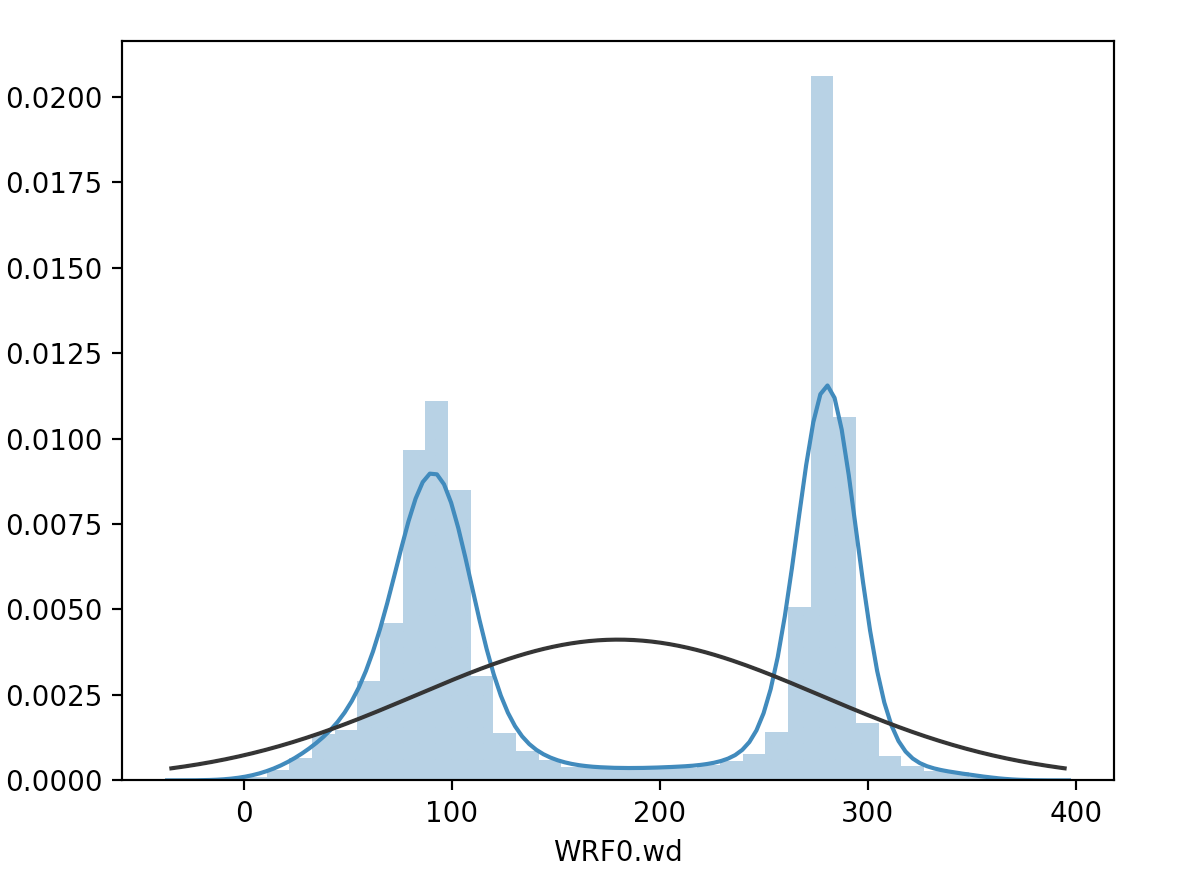


Figure 4. The Distribution of GFS0 Attributes





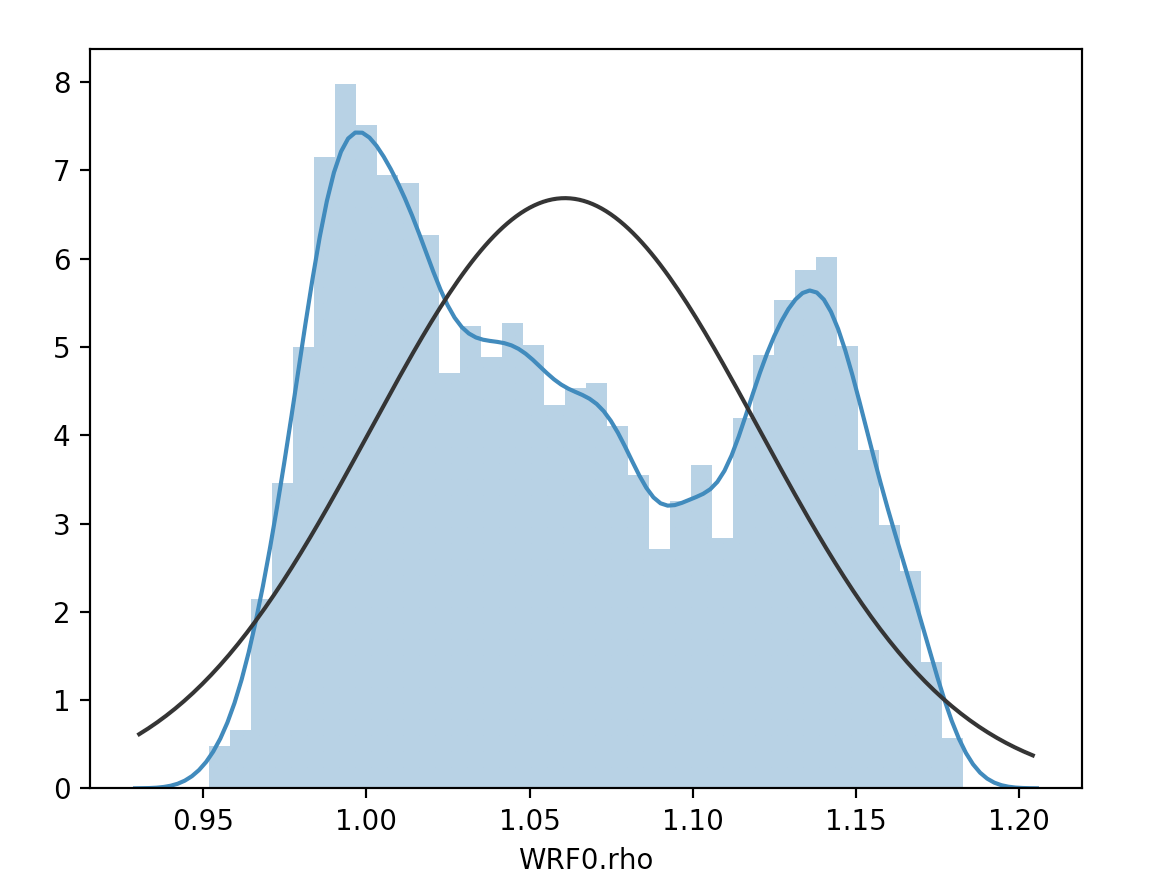
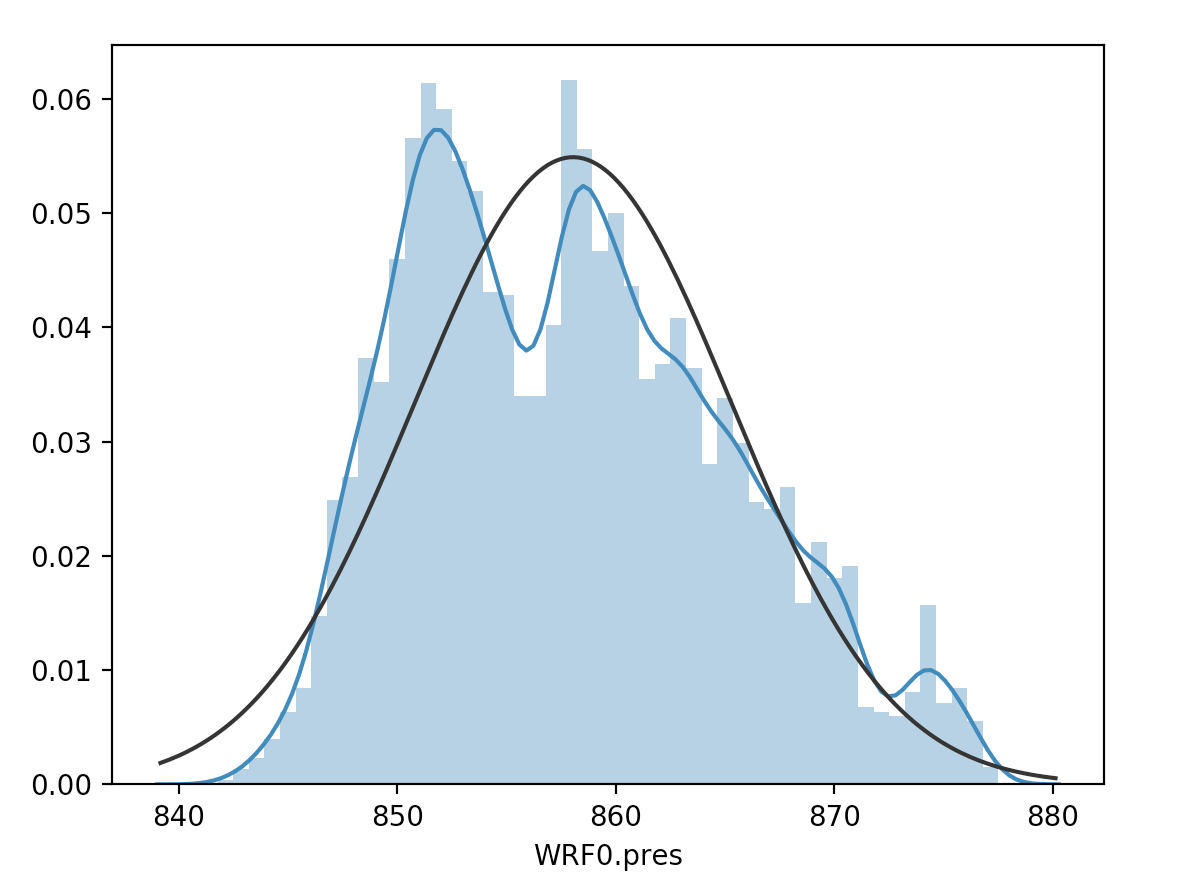


Figure 5. The Distribution of WRF0 Attributes

1.2.2 Heat Map of Correlation Among Data

Correlation is also important information on data exploration. We could see the heat map of one-dimensional linear correlation between attributes as Figure 6. The color of grid is lighter, the correlation relationship between two attributes is stronger.



Figure 6. Correlation of Attributes

Here, there is a table listing the value of target Y.ws\_tb correlation with other attributes.

Table 2. The Correlation between Other Attributes and Y.ws\_tb

|  |  |  |  |
| --- | --- | --- | --- |
| Y.ws\_tb | 1 | WRF0.rho | 0.092335 |
| Y.power\_tb | 0.952221 | GFS0.rho | 0.08942 |
| WRF0.ws | 0.676468 | EC0.pres | 0.087991 |
| EC0.ws | 0.668349 | WRF0.pres | 0.085148 |
| GFS0.ws | 0.585671 | EC0.rho | 0.076221 |
| GFS0.wd | 0.22851 | EC0.tmp | -0.070791 |
| WRF0.wd | 0.193127 | GFS0.tmp | -0.085529 |
| EC0.wd | 0.187612 | WRF0.tmp | -0.08935 |
| GFS0.pres | 0.099403 |  |  |

We could see from Table 2, final target Y.power\_tb has a strong proportional relationship with Y.ws\_tb. Besides, three measurements of wind speed data also have a strong relationship compared to other data. The correlation value of all temperature data nears to 0, which means they have a very small correlation with the target.

1.2.3 Scatter Plot of Wind Speed

We have known that wind speed data from three whether sources are important for the prediction. Now, we are going to see the scatter distribution of wind speed and target as the Figure 7.

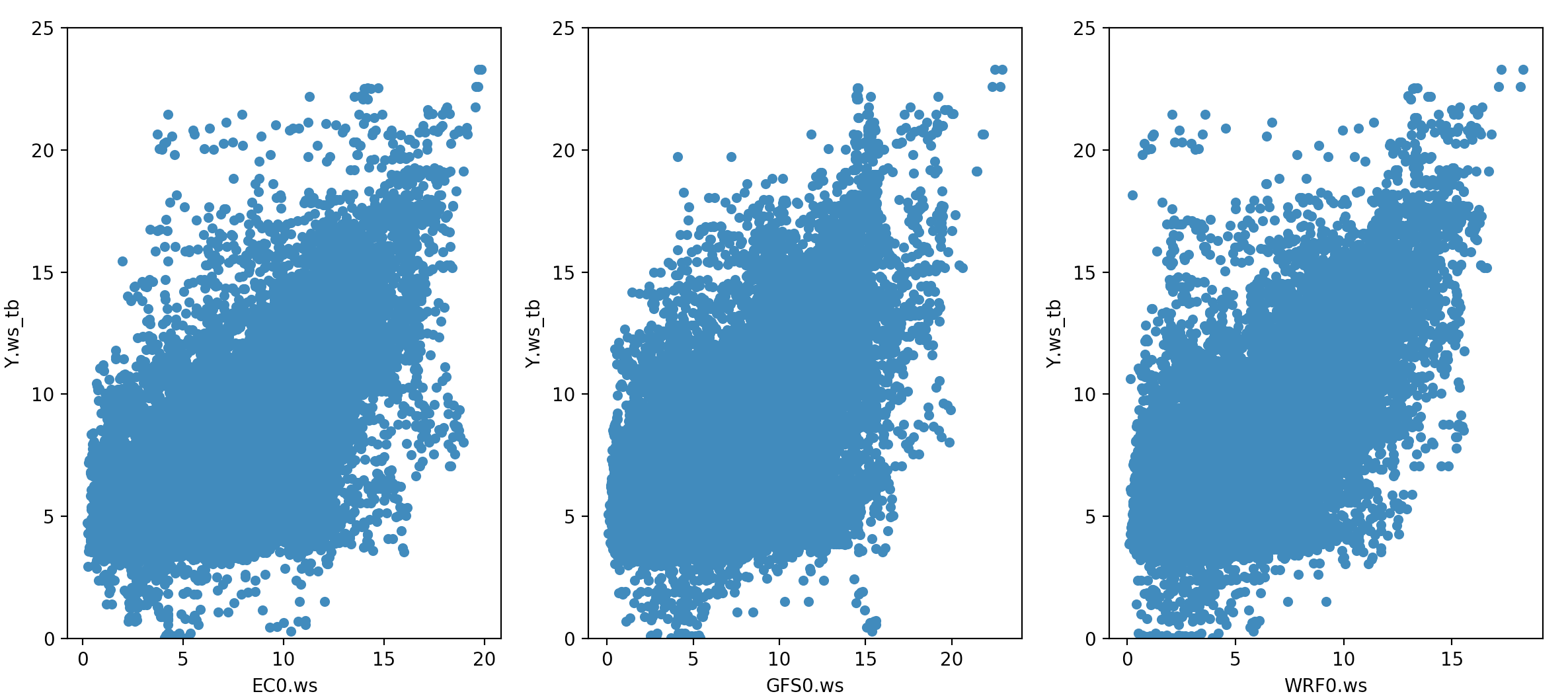


Figure 7. The Scatter Plot of Wind Speed and Y.ws\_tb

From the Figure 7, according to our common sense, we may infer there are some outliers. (the real wind speed is larger but the data from whether source is small, and vice versa). It is possible to find and delete this part of data on data pre-processing.

1.2.4 Scatter Plot of Wind Direction

Besides wind speed data, wind direction information is the second important. As from previous graph, we have known that the shape of the distribution of wind direction data is like two-humps. To be more specific, the wind direction is at range [0, 150] and [250, 360] has larger value of wind speed compared to other ranges.

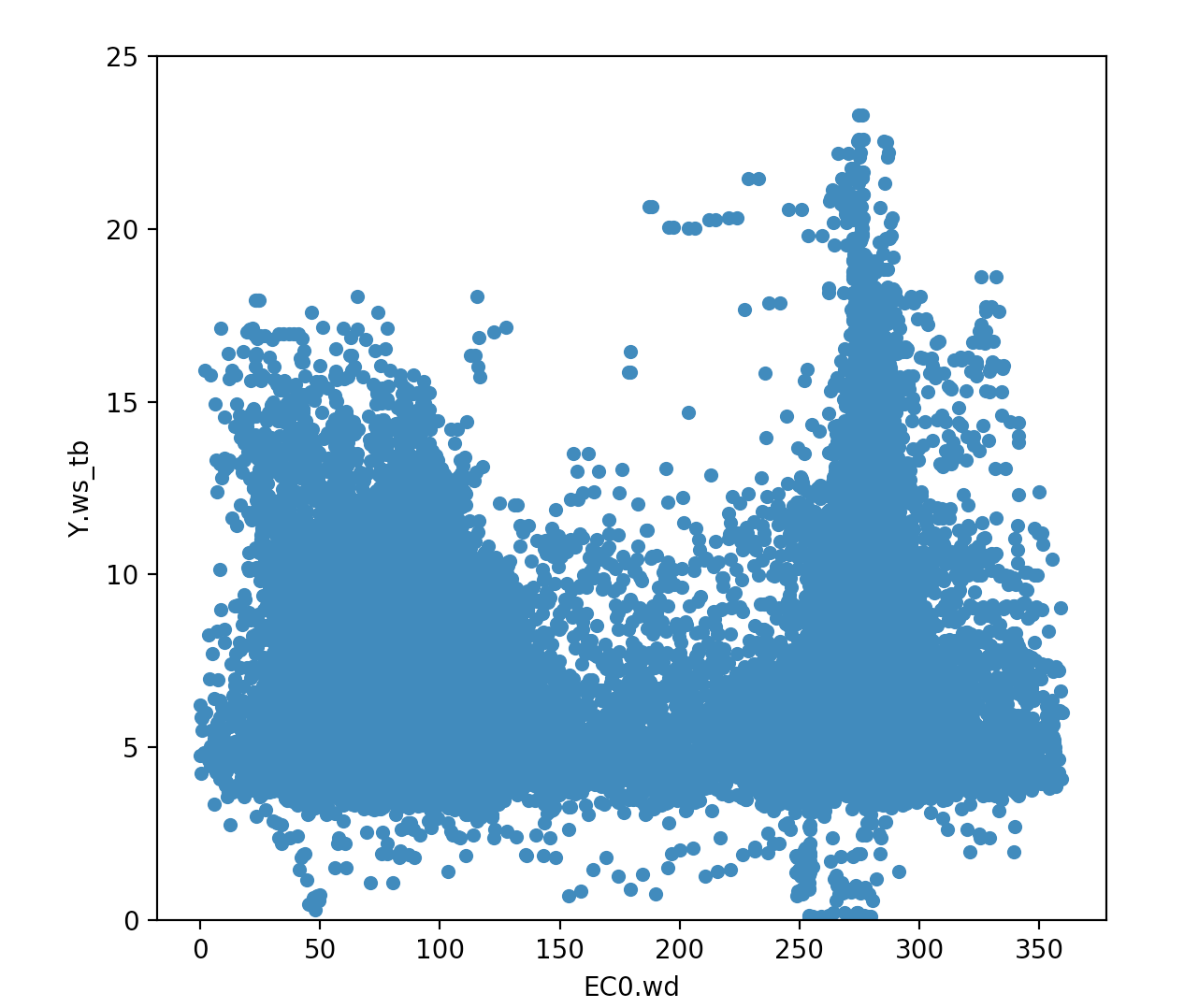
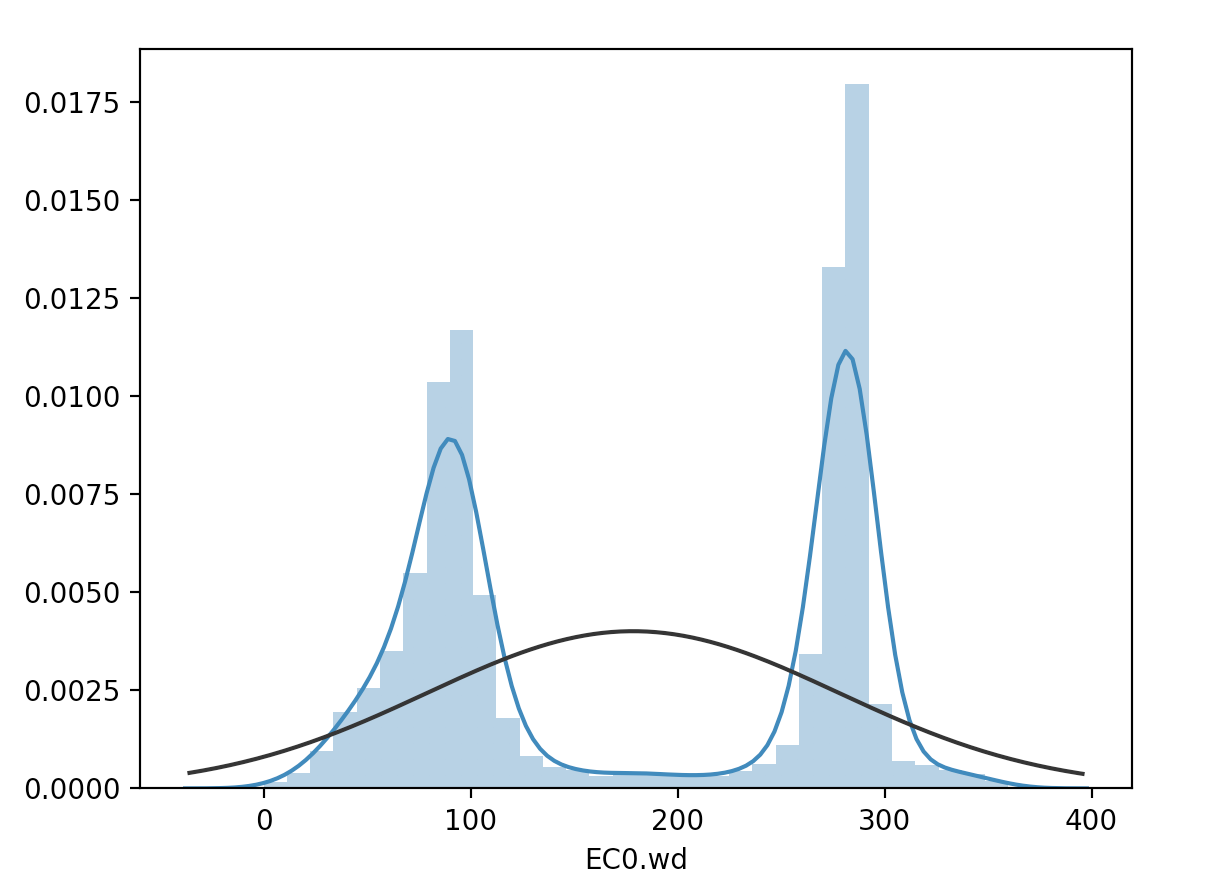


Figure 8. Scatter Plot of Wind Direction of EC0

Another two whether resources about the wind direction are similar as the Figure 9.

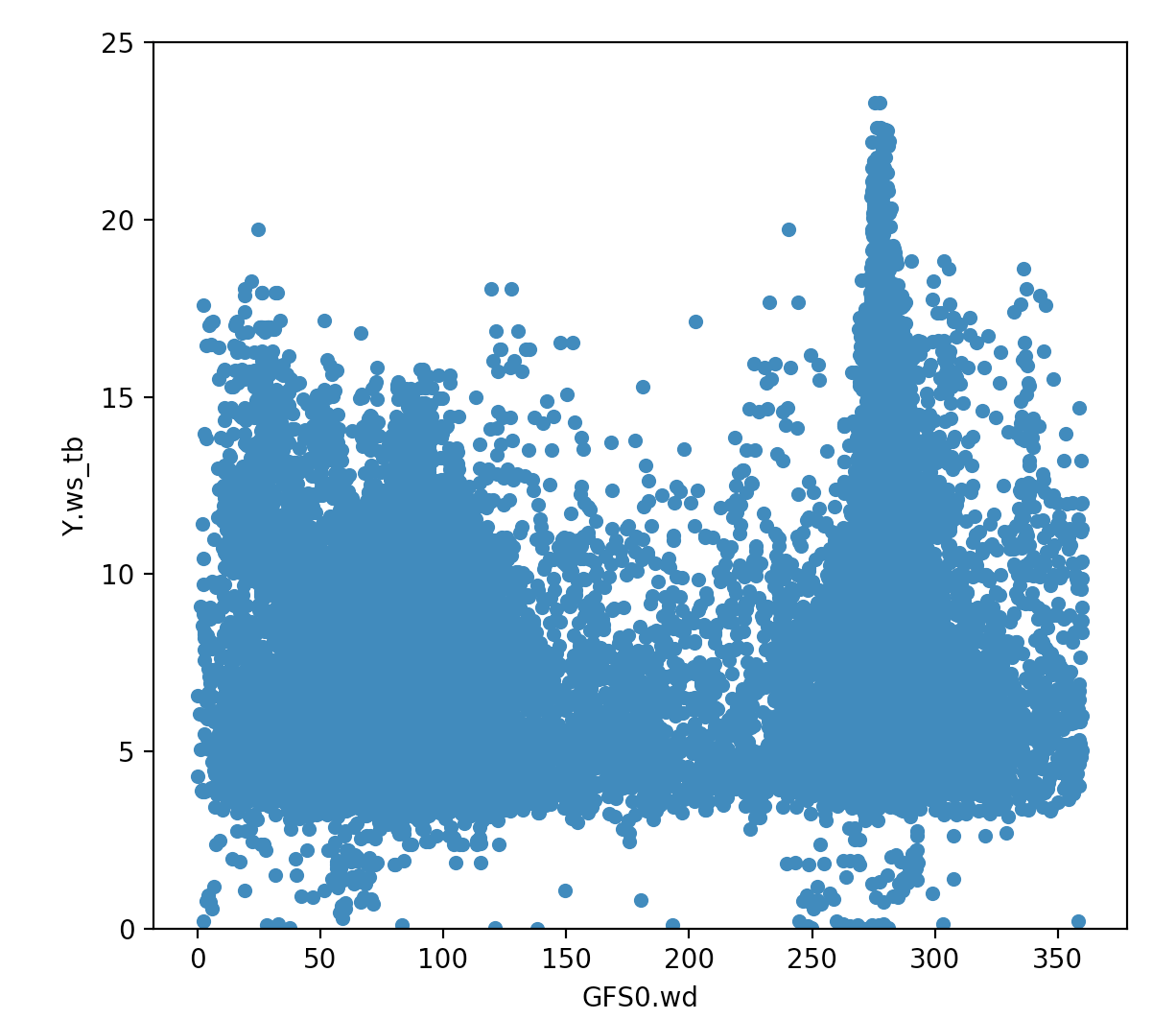
 

Figure 9. The Scatter Plot of Wind Direction of GFS0 and WRF0

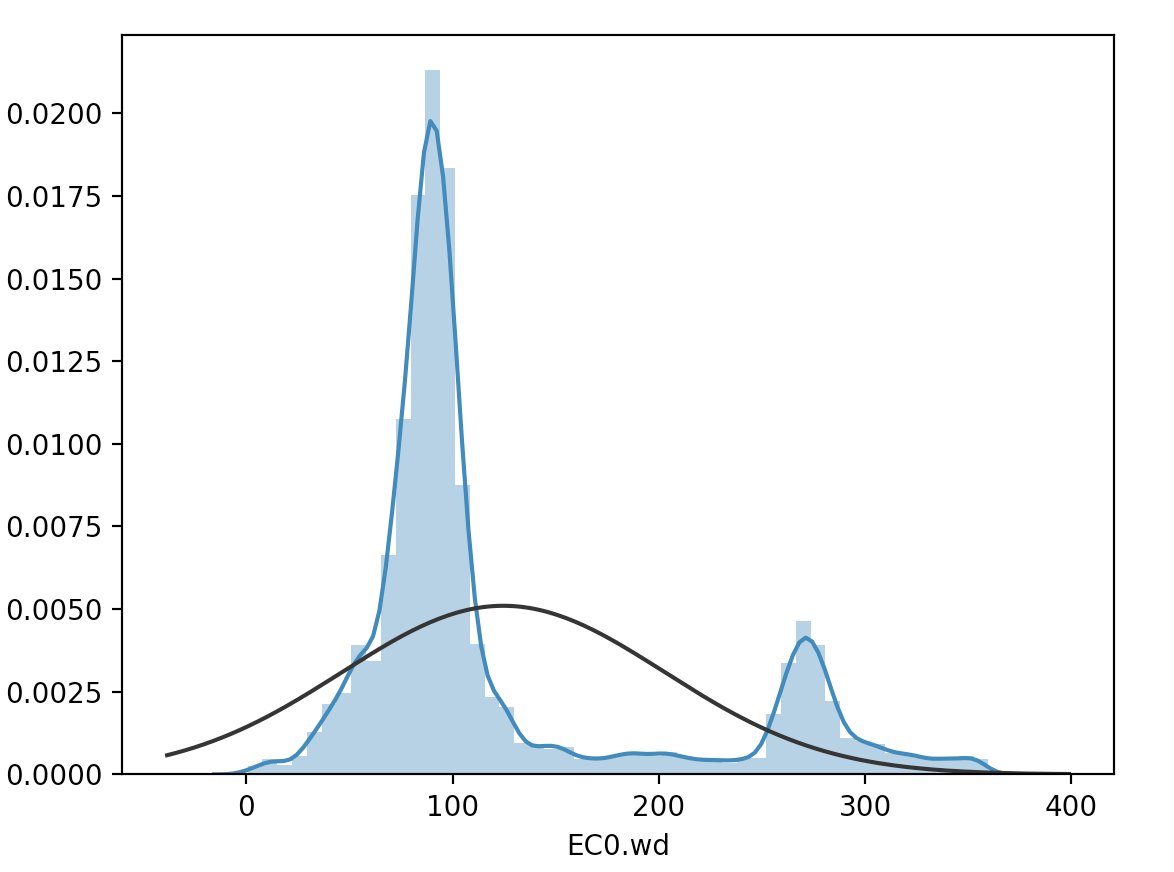
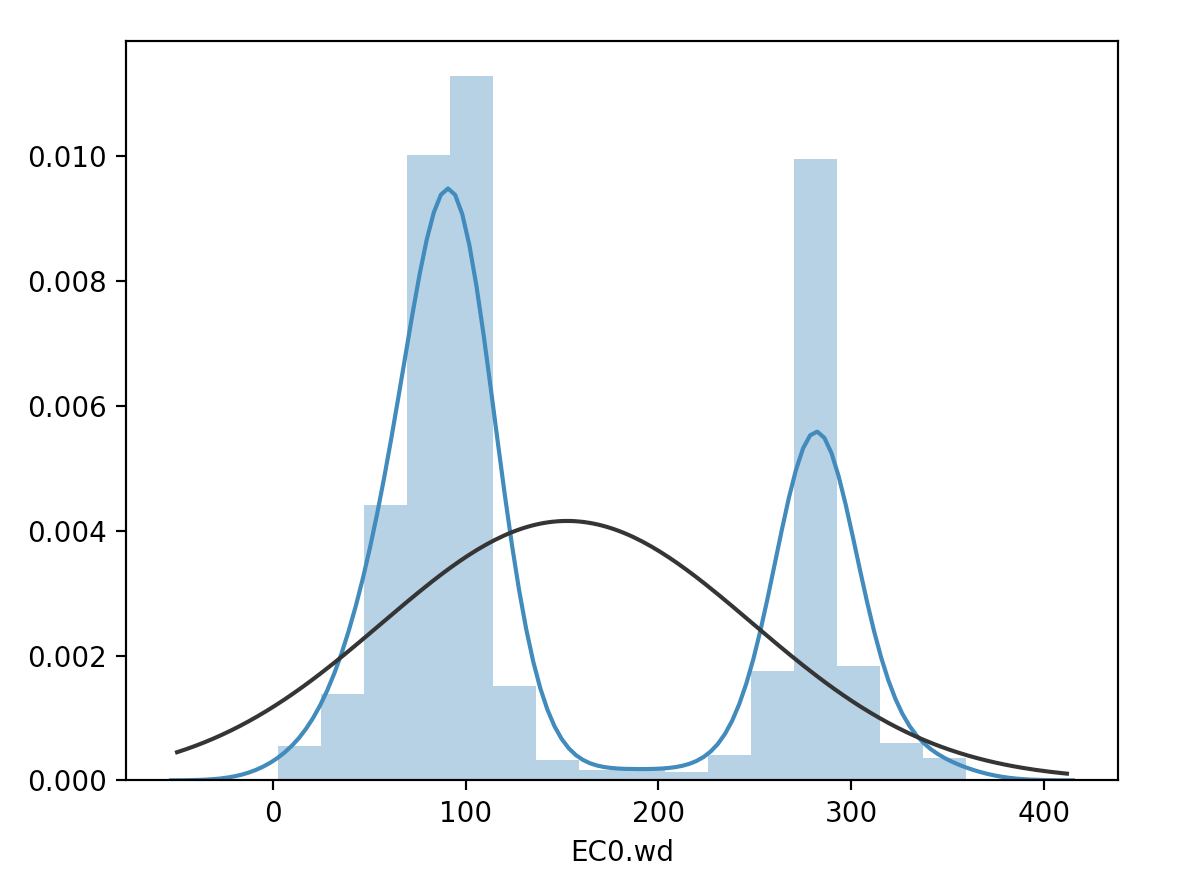
1.2.5 Inspection of Seasonal Wind Direciton

We divided the time into four seasons as following Table 3.

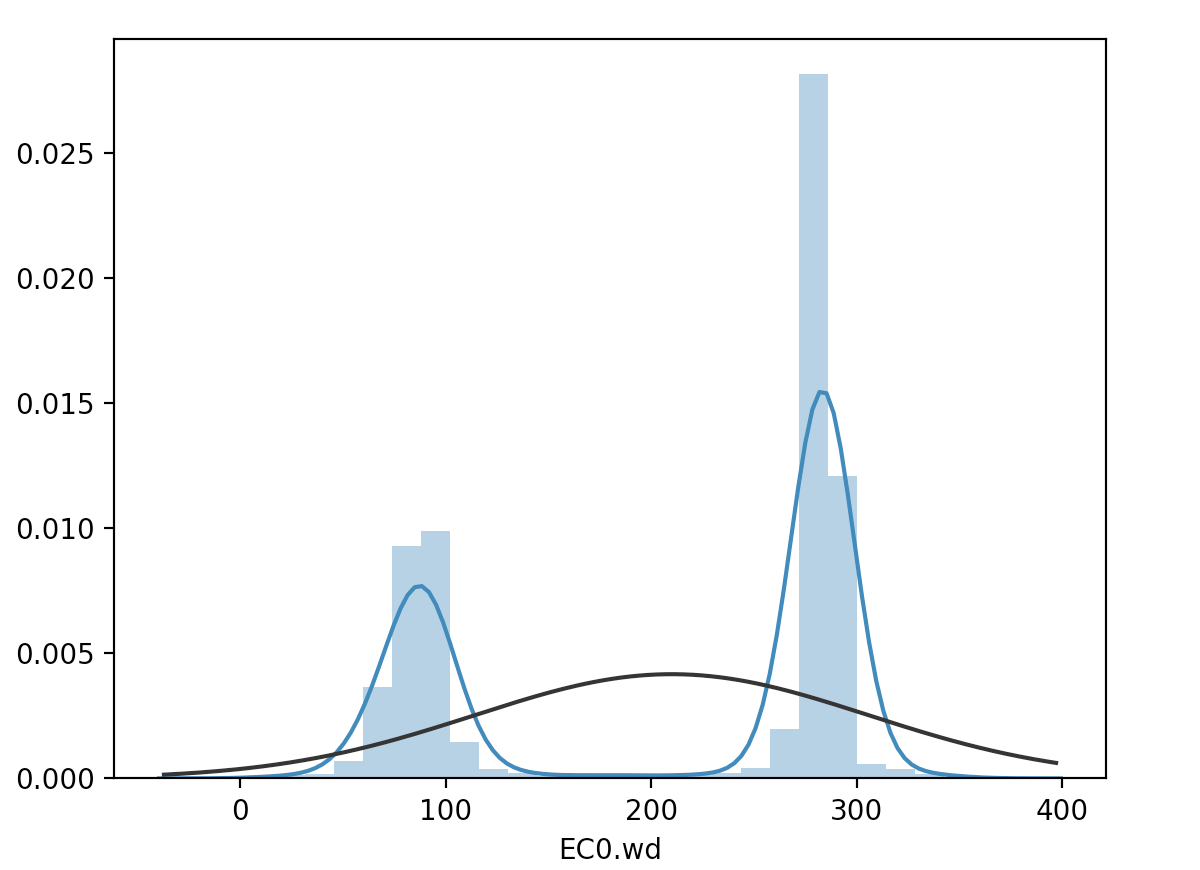
Table 3. Diding the Four Seasons

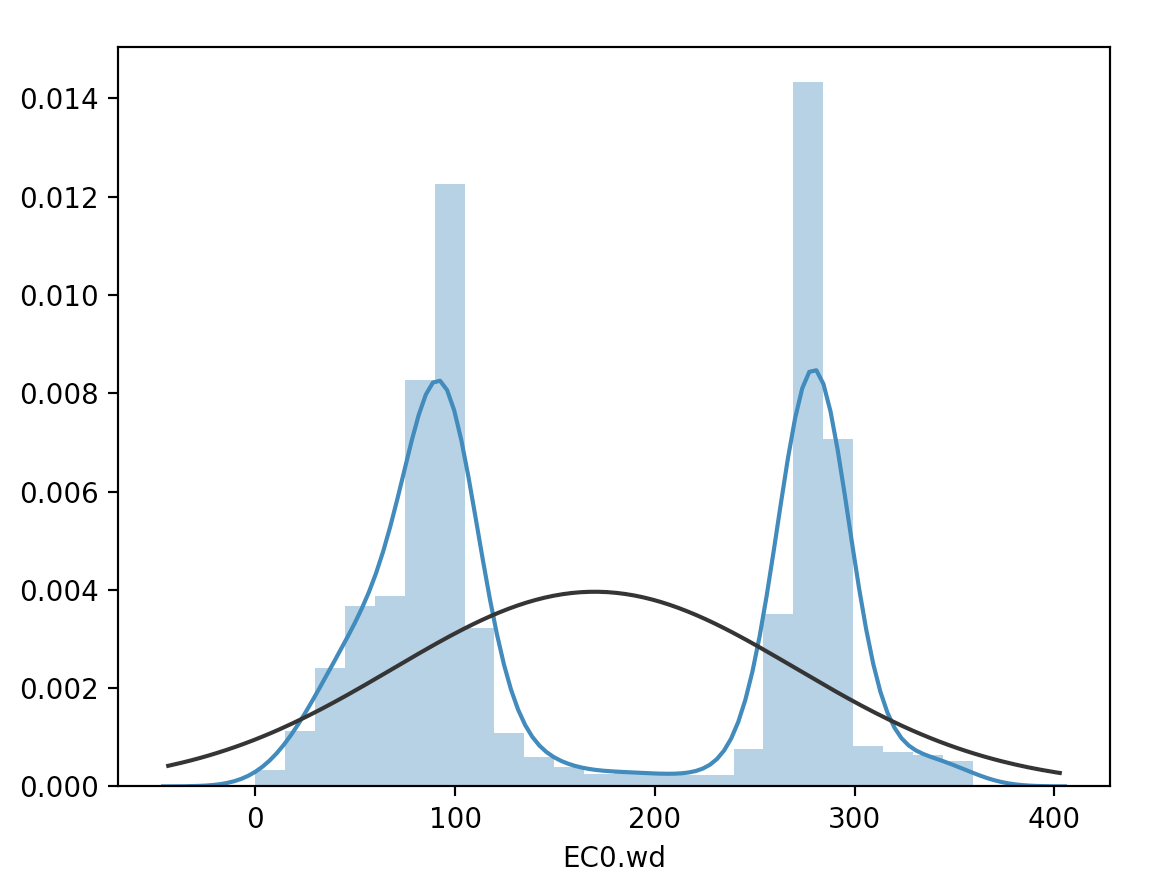
|  |  |  |  |
| --- | --- | --- | --- |
| Time | Season | Period | Set.id |
| 2017-07-01 – 2017-08-31 | Summer | 2 months | 0-61 |
| 2017-09-01 – 2017-10-31 | Autumn | 2 months | 62-121 |
| 2017-11-01 – 2018-2-28 | Winter | 4 months | 122-242 |
| 2018-03-01 – 2018-05-31 | Spring | 3 months | 243-334 |

We could see ftom the Figure 10, different season is different. there are the converse two-humps shapes in summer and winter respectivley. However, in Spring and Autumn, two humps have the same height. Hence, with the season changing, wind direction is differnet. Here, cosider adding season feature as a new attribute.



Time 17-07-01 – 17-08-31 Time 17-09-01 – 17-10-31





Time 17-11-01 – 18-02-28 Time 18-03-01 – 18-05-31

Figure 10. The Scatter Plot of EC0 Wind Direction in Four Seasons

1.2.6 Scatter and Box Plot of Atmospheric Pressure

Now, we see the relationship between atmospheric pressure and Y.ws\_tb as Figure 11. There is not a very clear linear relathionship between x asix and y asix varibales.

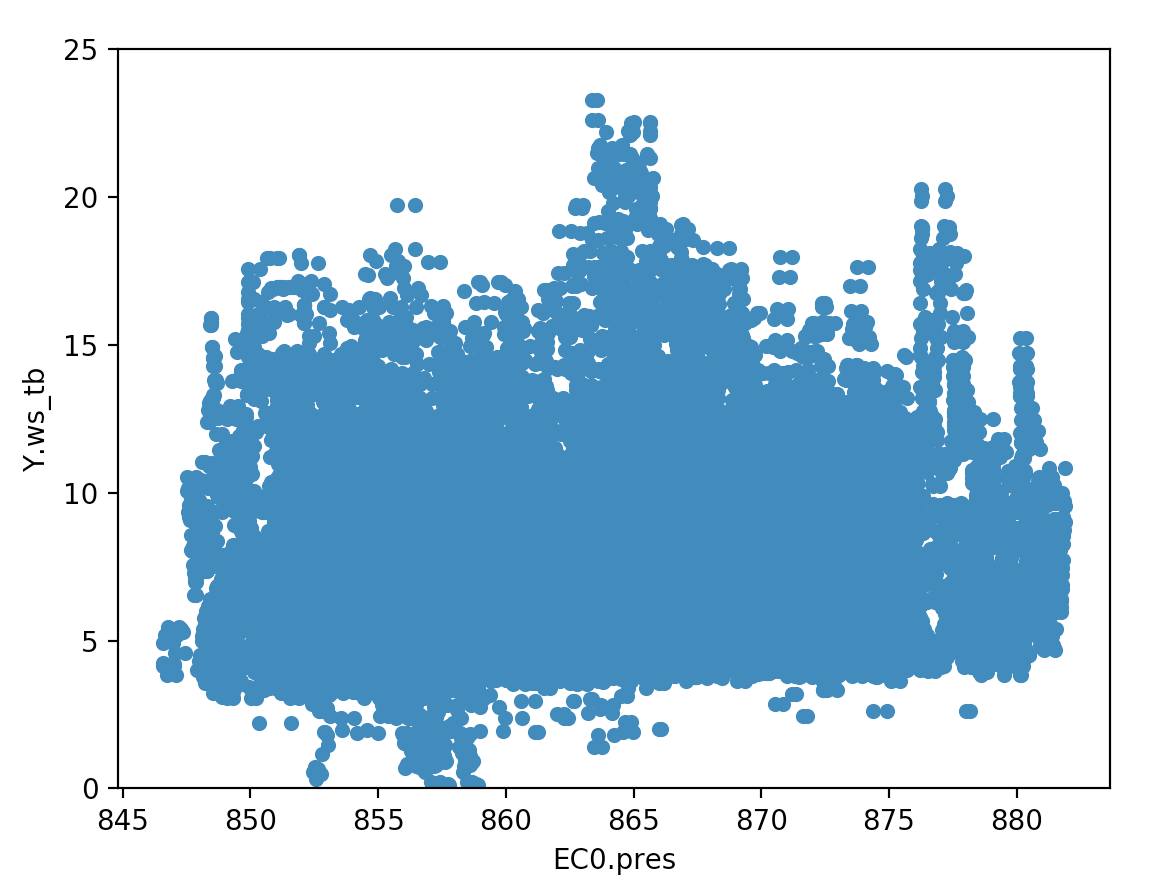


Figure 11. The Scatter Plot of EC0 Atmospheric Pressure and Y.ws\_tb

After rounding each atmospheric pressure value, draw the box plot as the Figure 12. We could see that every bin has nearly the same height. Hence, in this case, atmospheric pressure has a less strong relathionship with the target. This attribute needs to process and transform later.

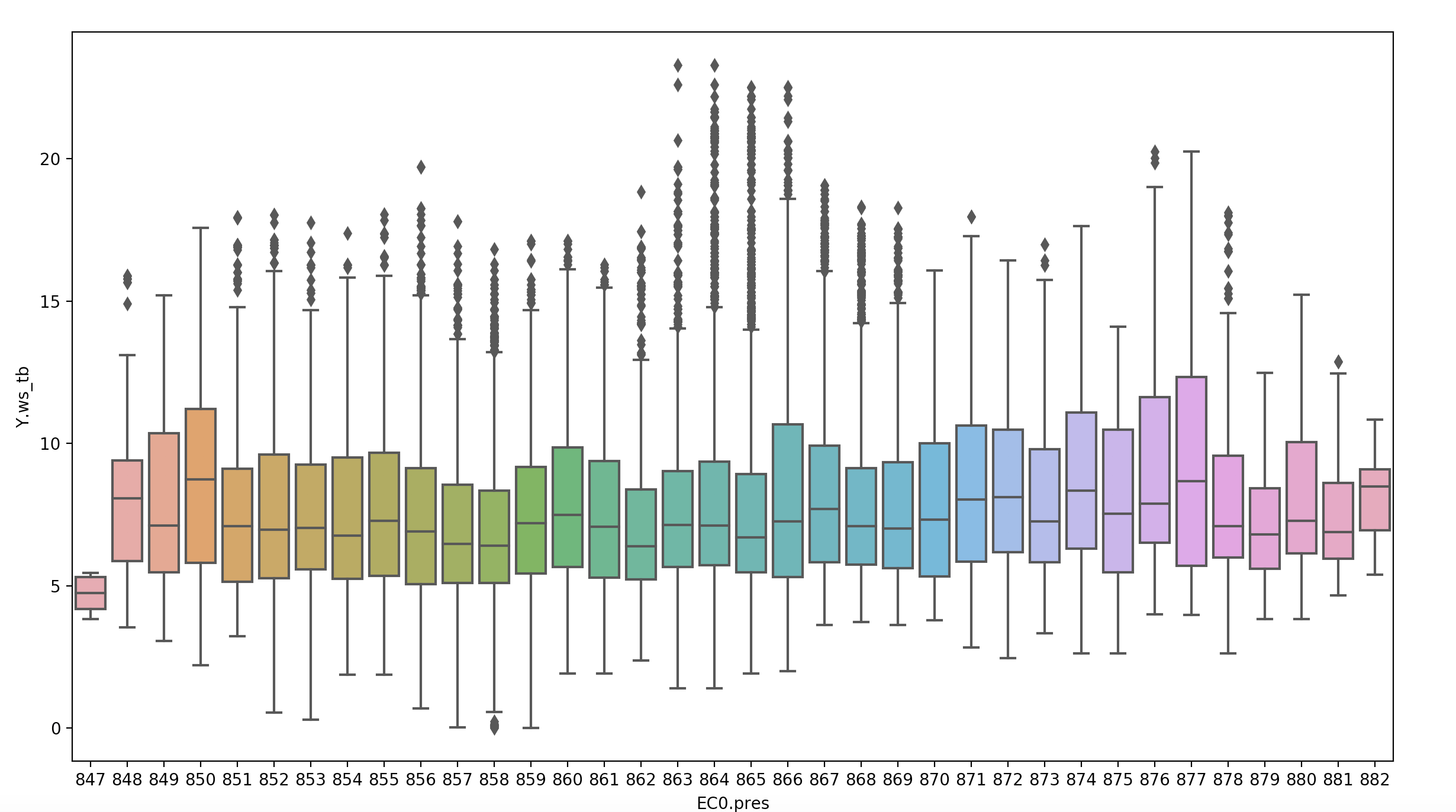


Figure 12. The Box Plot of EC0 Atmospheric Pressure

1.2.7 Scatter Plot of Wind Density

The relathionship between the wind density and Y.ws\_tb is very similar with that of atmospheric pressure and Y.ws\_tb. It could not be seen any distribution. If this attribute is used directly, which may not have a good effect on prediciton.

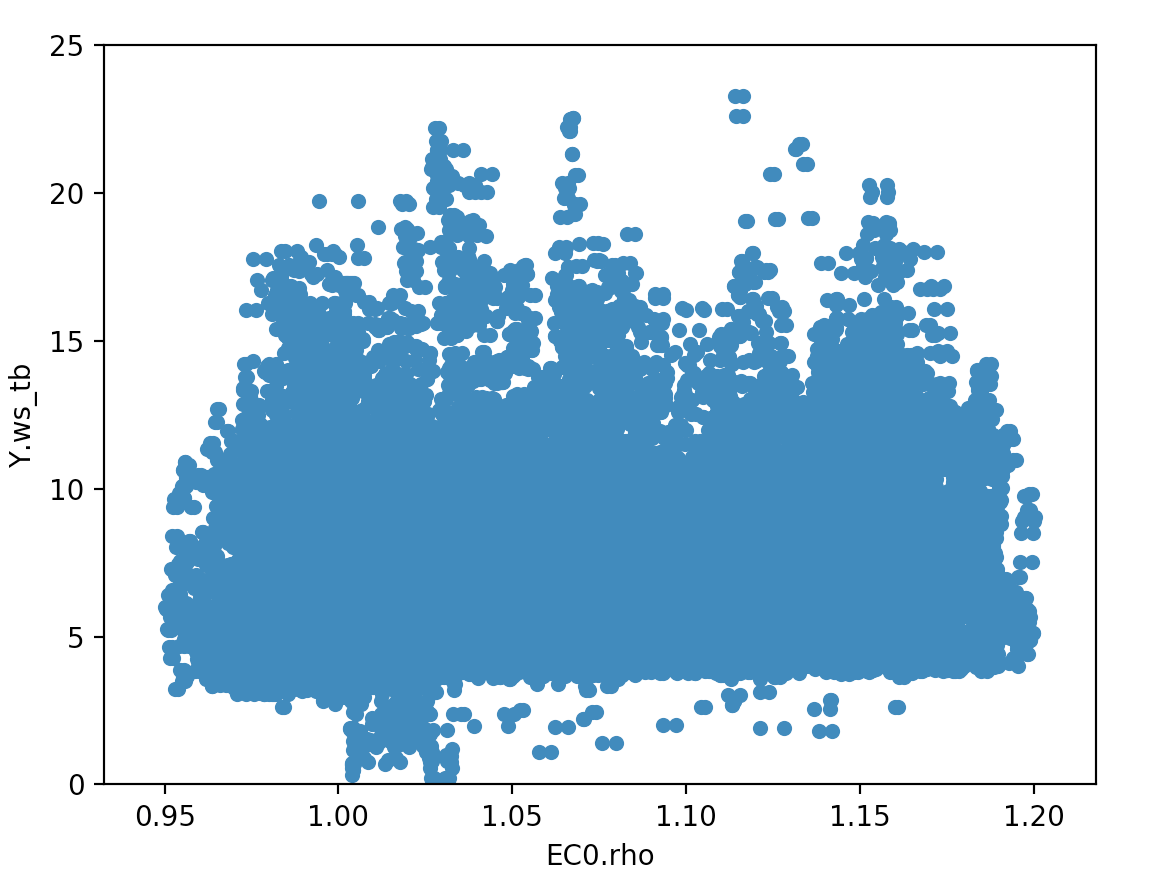


Figure 13. The Scatter Plot between EC0 Wind Density and Y.ws\_tb

1.2.8 Scatter Plot and Honeycomb Graph of Temperture

The correlation between the temperture data of EC0 and the taeget is also samll. It is easily find that so far the plot is the same as former two attributes: atmospheric pressure and wind density.

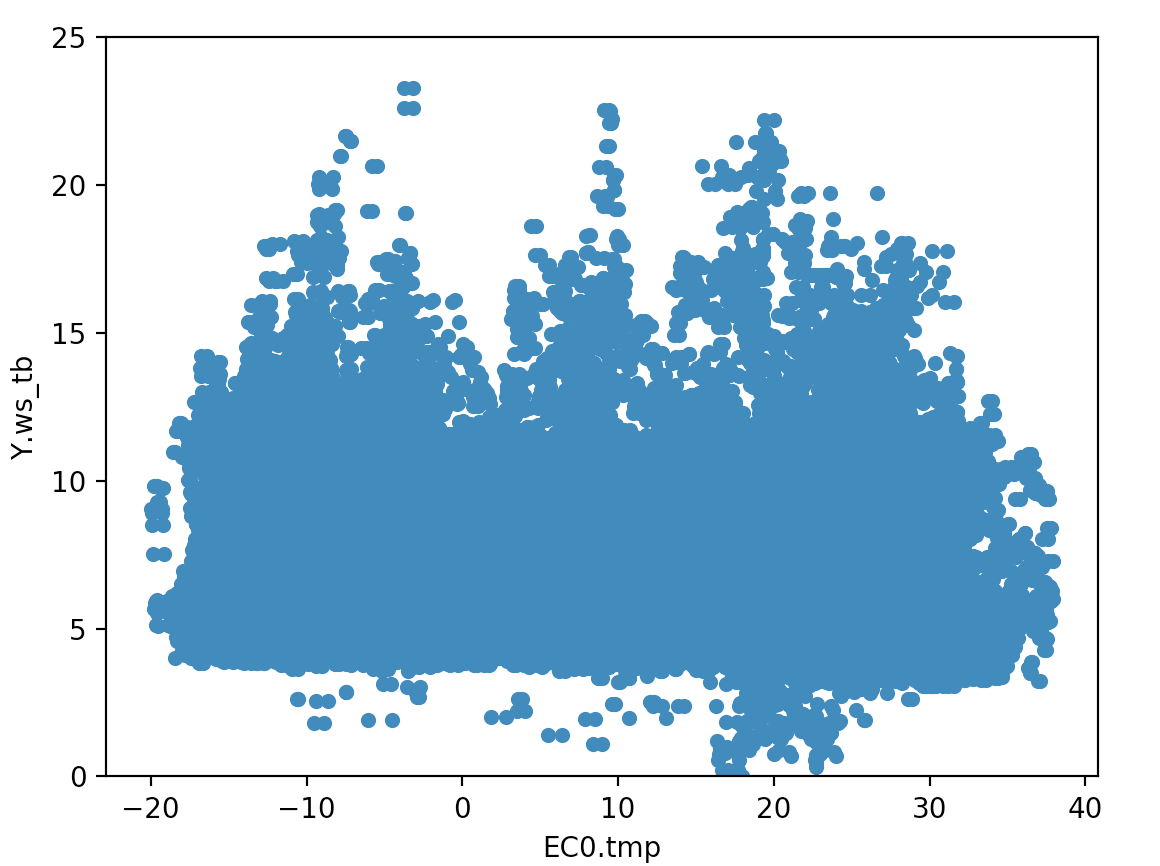


Figure 14. The Scatter Plot between EC0 Tempreture and Y.ws\_tb

To see the frequency of data, we draw honeycomb graph as Figure 15.

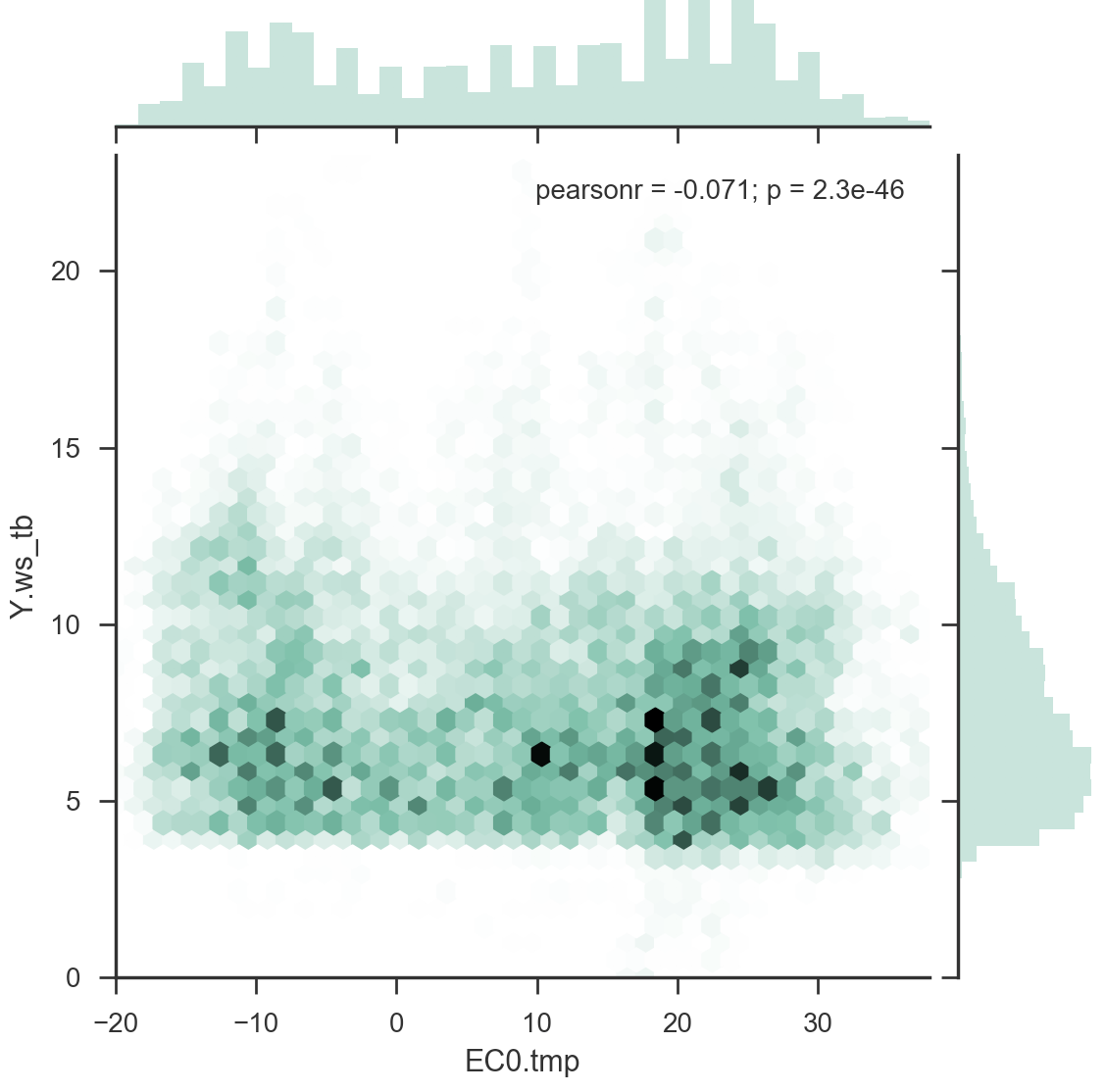


Figure 15. The honeycomb between EC0 Tempreture and Y.ws\_tb

1. **Data Pre-processing**

2.1 Smoothing the Target Y.ws\_tb on Training Data

Becaouse the data Y.ws\_tb have a significant fluctuation sometimes, we could consider smoothing it. For a single Y.ws\_tb value, we use the previous three vlaues and after two values (also include itself). There are two specilal conditions: when itself is nan value, we don’t do smooth and just keep nan. And when there is a nan value except itself, nan value is skipped. There are examples at Table 4.

Table 4. Examples of Smoothing Data

|  |  |  |
| --- | --- | --- |
| Y.ws\_tb value | Value Group | Smooth Value |
| 4 | [1,2,3,4,5,6] | (1+2+3+4+5+6)/6 |
| Nan | [1,2,3, nan,5,6] | Nan |
| 4 | [1, nan,3,4,5,6] | (1+3+4+5+6)/5 |

Especially, the first three and last two values in one set would keep the same. After comparing the results of non-smooth run and smooth run, we found that method of smoothing data on Y.ws\_tb is not very effective.

* 1. Adding Season and Month Attributes

From X\_basic.time, extracte the month and add it as a new feature. Besides, we can also use month information to divide up season and create another new feauture. The mapping of season as Table 5.

Table 5. Mapping of Season

|  |  |
| --- | --- |
| Season | Month |
| Spring | March to May |
| Summer | June to August |
| Autumn | September to November |
| Winter | December to February |

* 1. Adding Time Period Attributes

From X\_basic.hour, we could get the distance between current time and start time. Map the time into morning, afternoon and evening. The mapping of time period as Table 6.

Table 6. Mapping of Time Perood

|  |  |  |
| --- | --- | --- |
| Time Period | Time | Hours |
| Morning | 5AM – 12PM | 8 |
| Afternoon | 1PM – 8PM | 8 |
| Evening | 9PM – 4AM | 8 |

* 1. Discretization on Wind Direction

The wind direction ranges from 0 to 360, considering the separation of data, it can be discretized. Set 60 as an interval. [0,60], [60,120]…

* 1. Discretization on Temperature

The temperature mainly ranges from -10 to 30. We could also discretize it into several intervals as Table 7.

Table 7. Discretization on Temperature

|  |  |
| --- | --- |
| Interval | Temperature |
| 1 | [-99, -10] |
| 2 | [-10, 0] |
| 3 | [0, 10] |
| 4 | [10, 20] |
| 5 | [20, 30] |
| 6 | [30, 99] |

2.6 Data Dictionaty

Table 8. Data Dictionaty

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | Meaning | Type | Comments |
| X\_basic.time | 记录时间 | Datetime | Include year, month and day |
| X\_basic.horizon | 记录数据距离该时间距离 | Int | 10 as unit distance and range from [0，2880] |
| i.set | 数据组号 | Int | Train data include 394 sets and test data include 30 sets |
| EC0.ws | 风速 | Int |  |
| EC0.wd | 风角度 | Int |  |
| EC0.tmp | 温度 | Int | Discretize (6 categorires) and one-hot encoding |
| EC0.pres | 大气压 | Int |  |
| EC0.rho | 风密度 | Int |  |
| GFS0.ws | 风速 | Int |  |
| GFS0.wd | 风角度 | Int | Discretize (6 categories) and one-hot encoding |
| GFS0.tmp | 温度 | Int |  |
| GFS0.pres | 大气压 | Int |  |
| GFS0.rho | 风密度 | Int |  |
| WRF0.ws | 风速 | Int |  |
| WRF0.wd | 风角度 | Int |  |
| WRF0.tmp | 温度 | Int |  |
| WRF0.pres | 大气压 | Int |  |
| WRF0.rho | 风密度 | Int |  |
| New Attributes | | | |
| Season | 季节 | Category | From X\_basic.time extracting the month, [3,5] spring [6,8] summer, [9,11] autumn  [12, 2] winter. Then one-hot encoding |
| Month | 月份 | Category | From X\_basic.time extracting the month and one-hot encoding [1,12] |
| Time | 早上/下午/晚上 | Category | From X\_basic.hour getting the time distance, and then ont-hot encoding |

1. **Machine Learning Models**

3.1 The model with single whether source

Firstly, use the data from single source. There are total 5-dimension data from each weather source, mainly including wind speed, wind direction, temperature, pressure and density.

Wind Speed Prediction Results

Prediction Model

GradientBoostingRegressor

5-Dimension from Single Weather Source

Figure 16. Flow Diagram of Single Data Source

Table 9. STD Result for 10 Wind mills Using Three Weather Sources

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Data Set | EC0 | | GFS0 | | WRF0 | |
| Std train | Std test | Std train | Std test | Std train | Std test |
| 1 | 2.65888 | 2.61469 | 2.84731 | 2.74096 | 2.68369 | 2.76114 |
| 2 | 2.24224 | 2.45415 | 2.41702 | 2.59004 | 2.23126 | 2.54029 |
| 3 | 3.09333 | 2.71124 | 3.19940 | 2.80567 | 3.07549 | 2.82094 |
| 4 | 2.43232 | 2.56252 | 2.63584 | 2.74444 | 2.47656 | 2.71255 |
| 5 | 2.35299 | 2.60253 | 2.53346 | 2.74507 | 2.39619 | 2.74968 |
| 6 | 2.31905 | 2.64059 | 2.52555 | 2.73485 | 2.33062 | 2.74741 |
| 7 | 2.24485 | 2.47723 | 2.42020 | 2.59897 | 2.23666 | 2.58762 |
| 8 | 3.13091 | 2.54465 | 3.24795 | 2.68348 | 3.11376 | 2.66557 |
| 9 | 2.22397 | 2.49783 | 2.41907 | 2.58120 | 2.22519 | 2.60607 |
| 10 | 2.31547 | 2.61416 | 2.50415 | 2.70170 | 2.30069 | 2.68919 |
| Mean | 2.50140 | 2.57195 | 2.67499 | 2.69263 | 2.50701 | 2.68804 |

* 1. Improved Model of Single Whether Source

Based on the model of 3.1, add a layer of linear regression model when we get three prediction values from three weather sources.

3 Wind Speed

Prediction Model

GradientBoostingRegressor

5-Dimension from Single Weather Source

Wind Speed Prediction Results

Prediction Model

Linear

Figure 17. Flow Diagram of Improved Single Data Source

Table 10. Result of RMSE and STD of Improved Single Model

|  |  |  |
| --- | --- | --- |
| Data Set | Std train | Std test |
| 1 | 2.52407 | 2.55647 |
| 2 | 2.08051 | 2.39791 |
| 3 | 2.91531 | 2.64167 |
| 4 | 2.29925 | 2.50565 |
| 5 | 2.20598 | 2.54313 |
| 6 | 2.16000 | 2.55706 |
| 7 | 2.07276 | 2.41596 |
| 8 | 2.96563 | 2.48432 |
| 9 | 2.06840 | 2.42775 |
| 10 | 2.14883 | 2.52988 |
| Mean | 2.344074 | 2.50598 |

3.3 The Model of Three Whether Sources

Combining data from three weather sources, there are total 15-dimension data. We could easily find that combining data from three weather sources is better than just data from one source in 3.1.

Wind Speed Prediction Results

Prediction Model

GradientBoostingRegressor

15-Dimension from Three Weather Sources

Figure 18. Flow Diagram of Data From Three Weather Sources

Table 11. Result of RMSE and STD for 10 Wind Mills

|  |  |  |
| --- | --- | --- |
| Data Set | Std train | Std test |
| 1 | 2.4605 | 2.56005 |
| 2 | 2.04119 | 2.37562 |
| 3 | 2.81782 | 2.64211 |
| 4 | 2.25347 | 2.52613 |
| 5 | 2.15531 | 2.5462 |
| 6 | 2.11767 | 2.56816 |
| 7 | 2.03477 | 2.41363 |
| 8 | 2.85585 | 2.52655 |
| 9 | 2.02735 | 2.45631 |
| 10 | 2.10104 | 2.52419 |
| Mean | 2.286497 | 2.513895 |

3.4 Inspection of Single Feature Engineering

Since now, we begin to do data per-processing and inspect the effect on perdiction. We set the seed(0) as the random seed in this part and keep the same model GBoosting.

Table 12. Inspection of Single Feature Engineering

|  |  |  |
| --- | --- | --- |
| Method | Mean STD | Change |
| No-preprocessing | 2.513895 | - |
| Discretization on Wind Direction | 2.482909 |  |
| Discretization on Temperature | 2.517689 |  |
| Season | 2.519909 |  |
| Month | 2.5175 |  |
| Time period | 2.502591 |  |

From the table above, we could see that discretization on wind direction and add a new feature time period (morning, afternoon and evening) reduce the STD. However, other methods seem to keep the same.

**4. Time sequence model**

* 1. Procedure of the Model

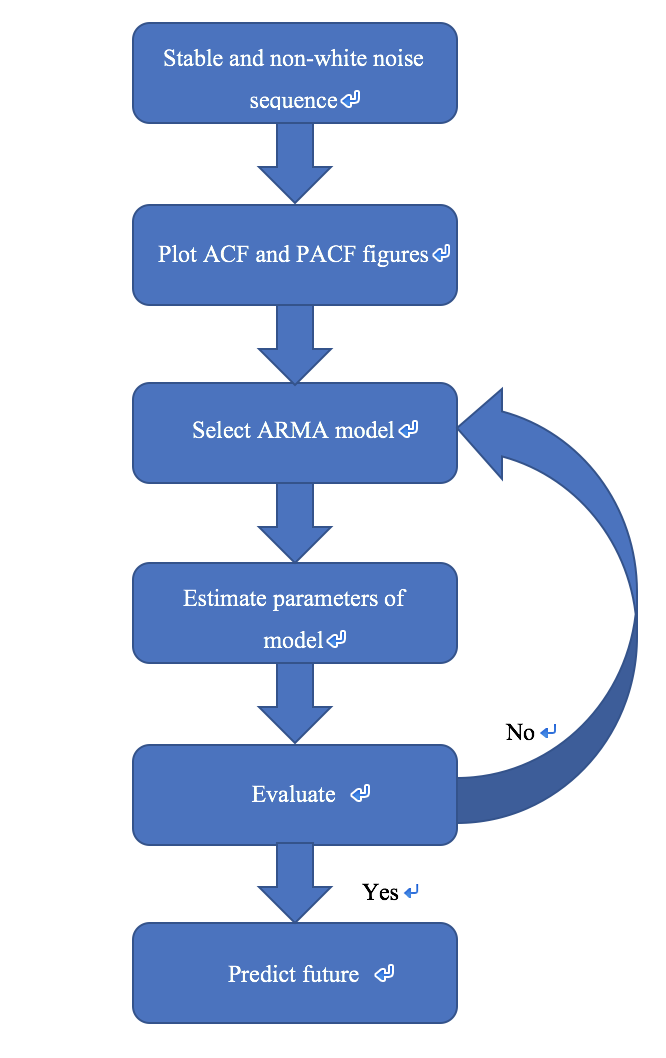


Figure 19. diagram of time sequence model

* 1. Experimental Data Set

To simplify experiment, choose a small part of data (2 days) and want to predict the future of data in 12h. For example, the range of train data is from 2017-07-12 00:00:00 to 2017-07-14 00:00:00. Prediction range is from 2017-07-14 00:10:00 to 2017-07-14 12:00:00.

* 1. Test Stable and Non-White Noise Sequence

For time sequence modeling, testing the given data whether it is stable and non-white noise is very important. As we known, only stable and non-white noise sequence has a prediction value, and could be used to build time series model.

4.3.1 Stable Sequence

The draw line graph of train data is as follows. From figure 20, we could see that train data fluctuate at some range. It seems that it is stable, but we cannot decide by intuition.



Figure 20. Time Sequence of Wind Speed

We use the ADF Test to decide whether it is stable sequence. The original assumption of ADF Test is non-stbale sequence. The result of test is as follows:

Test Statistic -2.859050

p-value 0.050321

Lags Used 0.000000

Number of Observations Used 288.000000

Critical Value (1%) -3.453262

Critical Value (5%) -2.871628

Critical Value (10%) -2.572146

P value is larger than 0.05, and we cannot reject the original assumption. Hence, the given sequence is not stable.

Luckly, we could transform data into a stable sequence with kinds of method. Here, we use finite differnece method to transform data. The figure of One order differnece is as figure 21.

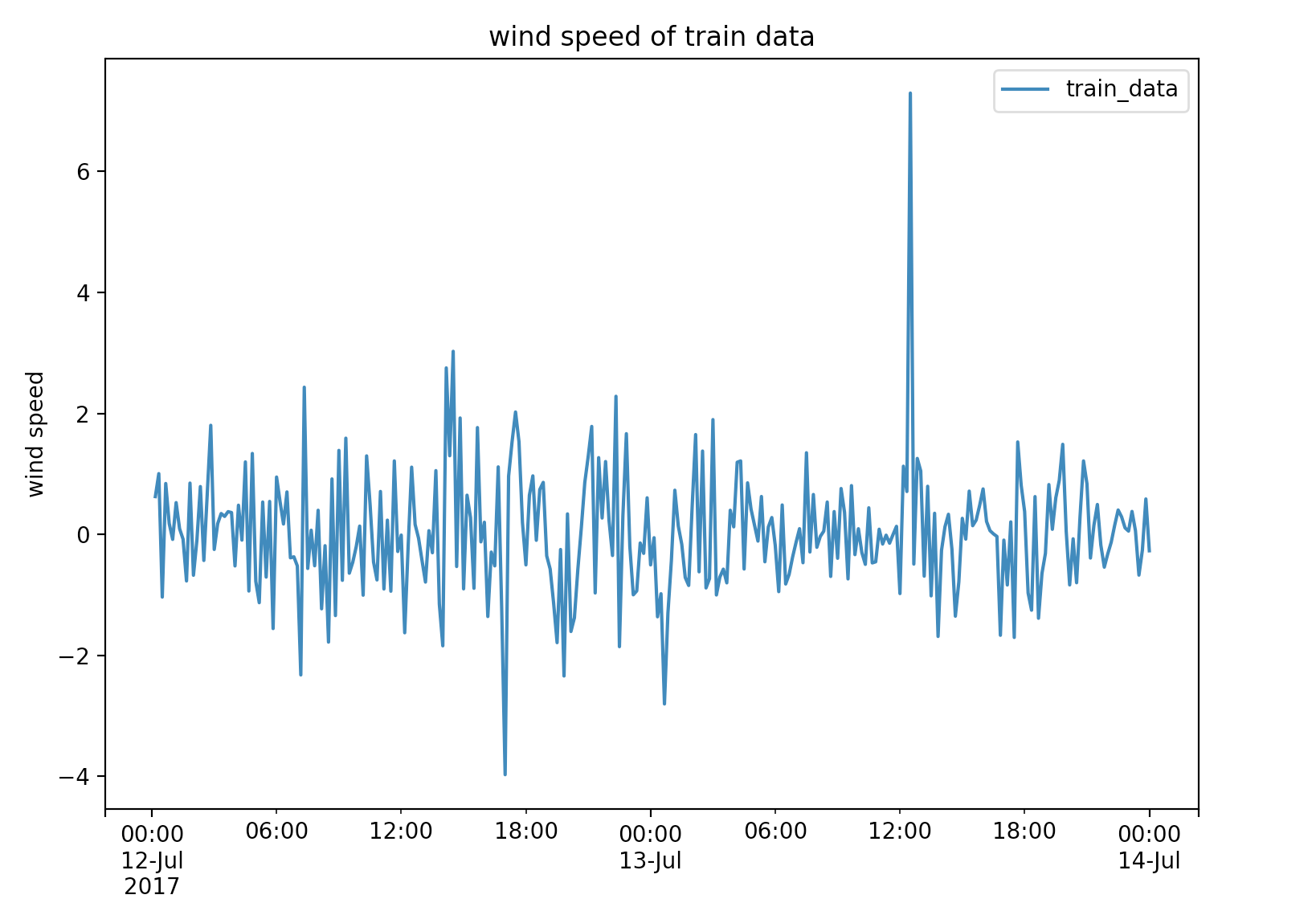


Figure 21. Time Sequence of Wind Speed after One Order Differnece

We do ADF Test again, and the result is as follows. Now, we could see the p value is smaller than 0.05, which means we should reject the original assumption. Hence, after the operation of one order differnece, data are stable.

Test Statistic -1.028706e+01

p-value 3.651489e-18

Lags Used 1.000000e+00

Number of Observations Used 2.860000e+02

Critical Value (1%) -3.453423e+00

Critical Value (5%) -2.871699e+00

Critical Value (10%) -2.572183e+00

4.3.2 Non-White Noise Sequence

White noise is a absolute random sequence, which is not useful for future prediction. If our data are white noise sequence, we cannot use sequence model to predict. There are two methods to check the given sequence whether is white noise or not.

The firest one is to use Ljung-Box Test (also called Q Test). We firstly assume that given sequence is white noise. Then do Ljung-Box Test, and the result is the below:

AC Q Prob(>Q)

lag

1.0 0.939026 257.485853 6.061082e-58

2.0 0.883519 486.225255 2.615336e-106

3.0 0.810967 679.614057 5.525753e-147

4.0 0.742777 842.417343 4.976517e-181

5.0 0.674429 977.110170 5.423620e-209

6.0 0.603530 1085.353910 3.076425e-231

7.0 0.538858 1171.948354 8.215275e-249

8.0 0.473553 1239.063790 3.474251e-262

9.0 0.420620 1292.202795 1.494486e-272

10.0 0.375211 1334.639205 1.277940e-280

11.0 0.331856 1367.954729 9.865031e-287

12.0 0.300177 1395.311524 1.425923e-291

13.0 0.268309 1417.247254 2.951989e-295

14.0 0.252689 1436.773988 1.963126e-298

15.0 0.246697 1455.453598 1.935617e-301

16.0 0.246156 1474.119451 1.876097e-304

17.0 0.252229 1493.789786 1.082178e-307

18.0 0.259717 1514.722280 3.279030e-311

19.0 0.256809 1535.264536 0.000000e+00

20.0 0.244519 1553.956922 0.000000e+00

Figure 22. The Result of Ljung-Box Test

We just need to observe the last column prob(>Q). Because of the value of this column is smaller than p 0.05, we reject the oraignal assumption(white noise ). Hence, the given sequence is not white noise.

The second method is to use autocorrelation analysis. We plot the autocorrelation figure of given data as figure 23.

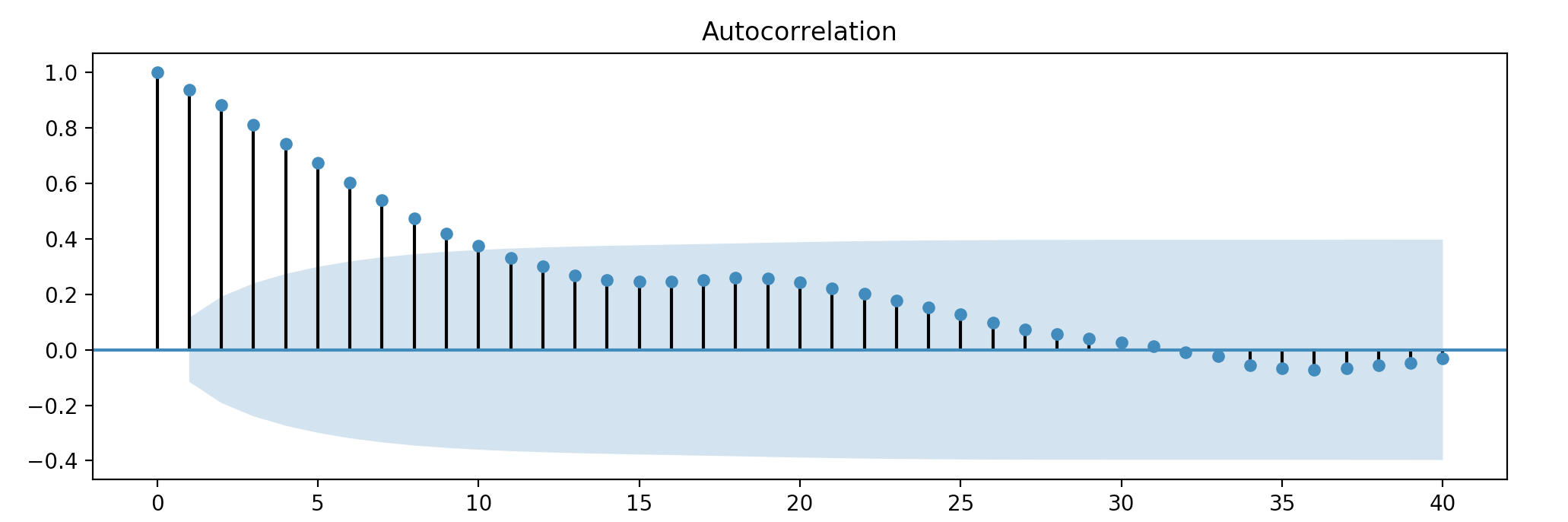


Figure 23. The Autocorrelation Graph of Given Data

From above figure, we observe that there are some correlation points not in the blue area except lag 0 point. Hence, we also get that given data are not white noise.

* 1. Plot ACF and PACF

ACF and PACF is important for us to choose which sequence model and determine the parameteres of the model. the figure is as follows:

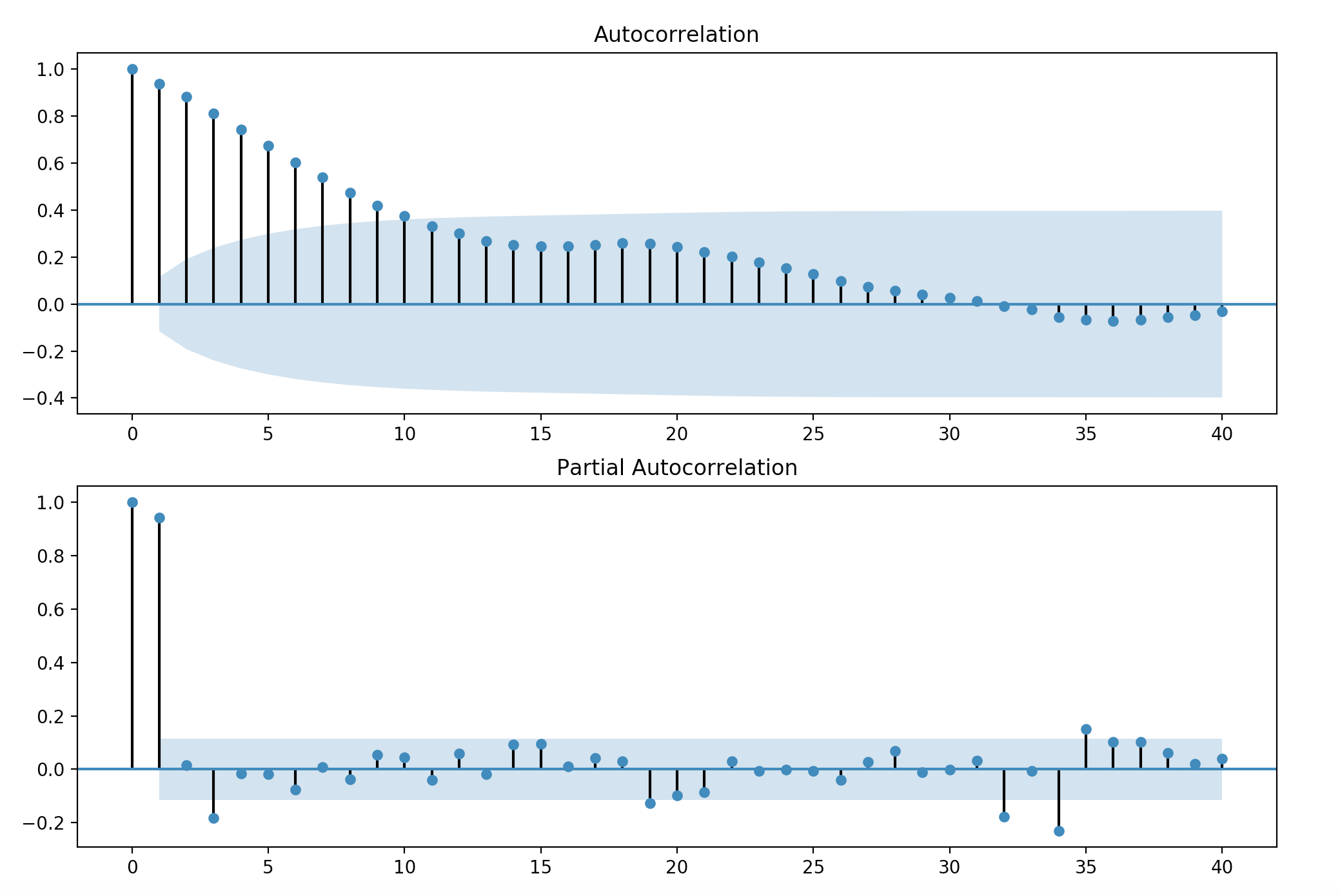


Figure 24. The ACF and PACF

From figure 24, we could see that for autocorrelation, it is hangover. However, for partial autocorrelation, it is truncation. Hence, we should choose AR model according to Figure 25.



Figure 25. Selection Method of Model

We also could determine the order of *p* in AR model. As usual, the value of order should not be too large, and in the figure, we choose 1 or 3.

* 1. Residua Check

After building AR model, we shold check residua. We mainly use two methods to check it, inclding ploting ACF and PACF of residua, and Ljung-Box Test.

From ACF and PACF, we could see that the seqeunce of residua is white noise basically.

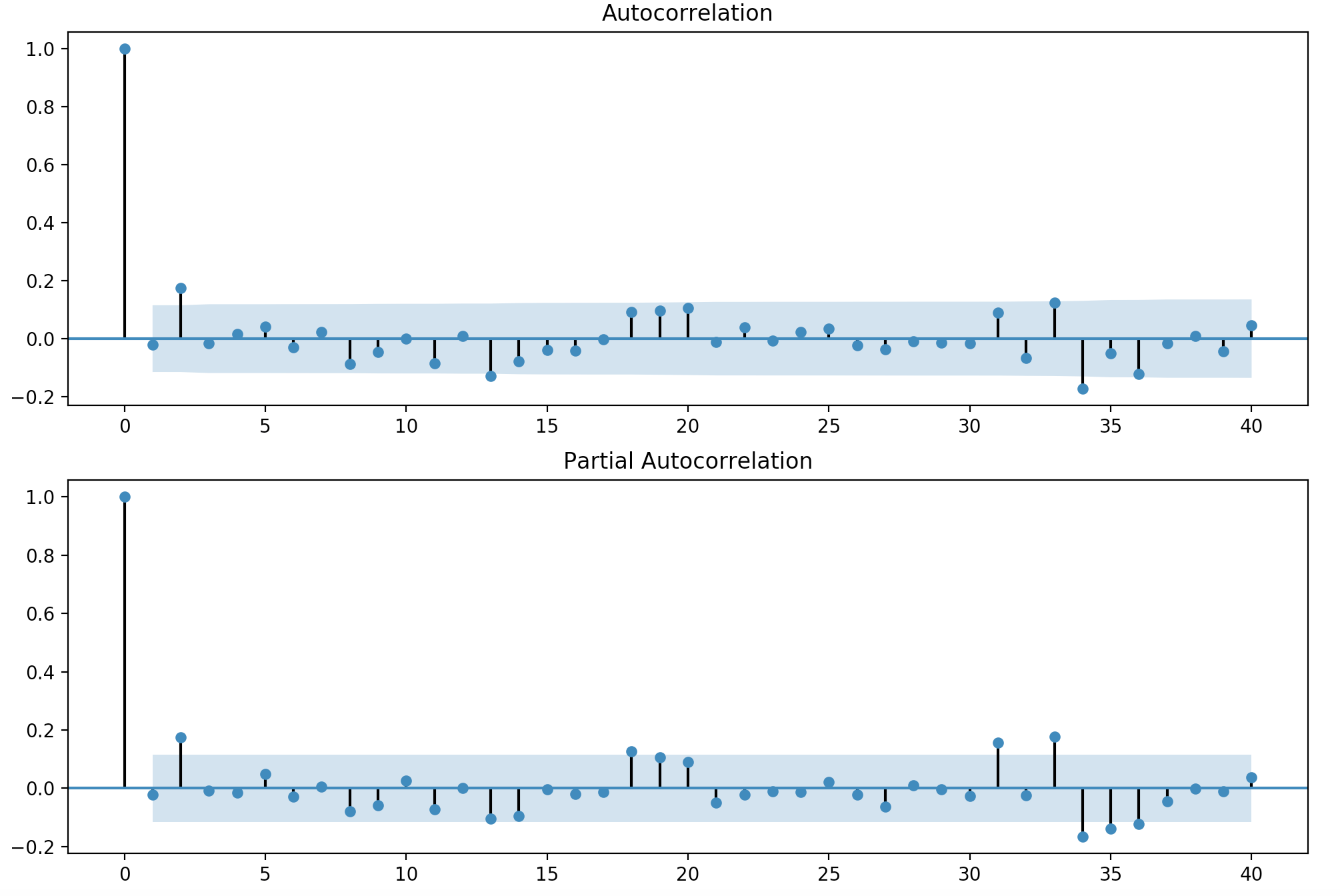


Figure 26. The ACF and PACF on Residua

We could get the same conclusion from the result of Ljung-Box Test. we could observe that all probability is larger than 0.05, which menas we cannot reject the origianl assumption. Hence, the residua is white noise and there are not more information that we could gain from the residua.

AC Q Prob(>Q)

lag

1.0 -0.005697 0.009478 0.922444

2.0 -0.010929 0.044481 0.978005

3.0 -0.005019 0.051887 0.996905

4.0 0.002696 0.054033 0.999642

5.0 0.059245 1.093413 0.954685

6.0 -0.000812 1.093609 0.981814

7.0 0.040089 1.572893 0.979678

8.0 -0.071381 3.097833 0.928069

9.0 -0.024921 3.284376 0.951946

10.0 0.029285 3.542887 0.965622

11.0 -0.041825 4.072087 0.967761

12.0 0.036169 4.469268 0.973406

13.0 -0.101392 7.601730 0.868545

14.0 -0.067640 9.000894 0.830993

15.0 -0.007693 9.019059 0.876519

16.0 -0.037513 9.452564 0.893613

17.0 -0.007804 9.471393 0.924303

18.0 0.086682 11.803152 0.857232

19.0 0.108974 15.502073 0.690198

20.0 0.088006 17.923464 0.592450

4.6 Predcition of Future

Using AR model with *p=1*, the figure of prediction is as follows. The blue line represents train data, green line is actual trend and orange line is our predicition.

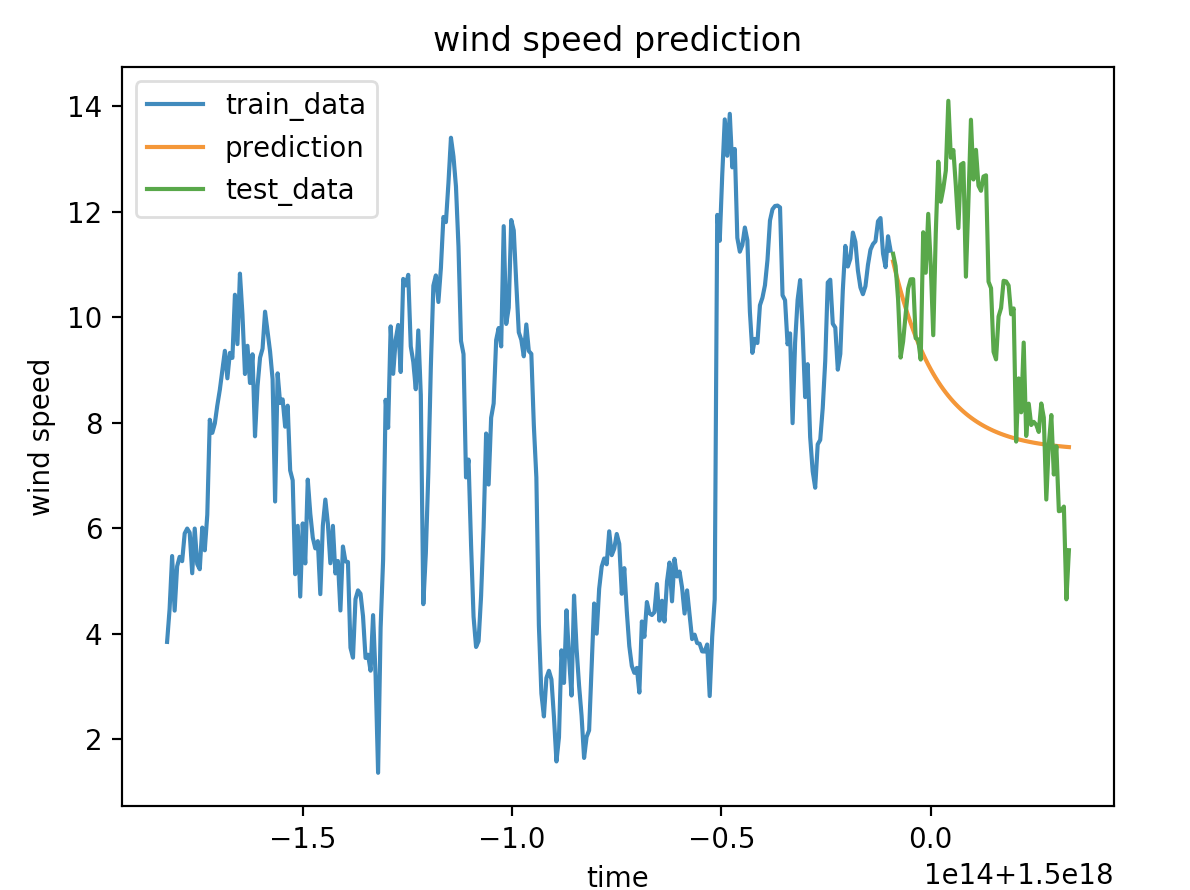


Figure 27. The Result of Presiciton with AR