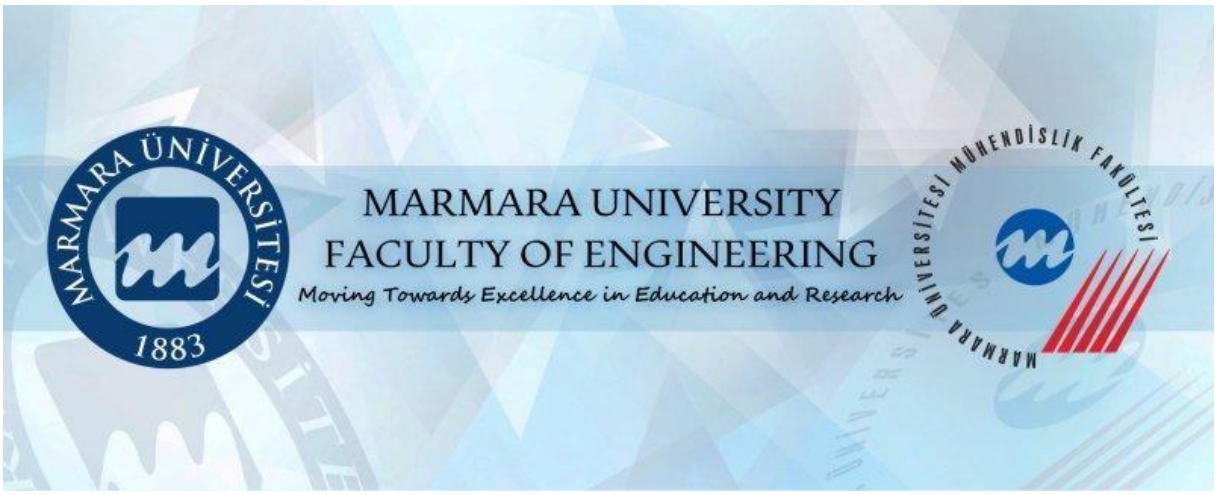


CSE4062 2020 SPRING PROJECT

ANALYSIS OF ELECTRIC PRODUCTION OF A WIND TURBINE

Delivery 3: Exploring Your Data Part II

Group: 8

Name	Department	ID Number	Mail Address
Furkan Can Ercan	Industrial Engineering	150316044	furkanercan98@gmail.com
Muhammed Avcı	Mechanical Engineering	150414007	muhammedavci96@gmail.com
Yasin Gök	Computer Engineering	150115058	ygok.96@gmail.com

3) Exploratory Data Analysis: Plotting Data

In the plotting phase of our project, we tried to demonstrate our attributes in a way, which will help us to enforce our previous understanding about our dataset. We started to plot our variables with a similar approach to our previous explore of the dataset. We first divided our attributes as numeric and categorical variables and then created our plots accordingly.

For our numerical attributes, we started with histograms and boxplots to see the general structure of the variable. To observe different details about the general structure of the regarding variable, we detailed our plots by adding some categorical sub divisions. For instance after plotting the histogram of the LV Active Power variable, we plotted the same histogram with day-night sub divisions. So that we were able to observe the possible differences between the distribution of the LV Active Power in daytime and night hours.

Additionally, we tried our best to discover relations between our variables. We especially focused on the linear model plots, which represent the linear relationship between two variables with a fitted regression line, to observe relationships between our variables. We tried to draw every possible combination to explore the dataset in a complete manner.

In the last part of our observations about the numerical attributes, we plotted the time series graphs to see general behavior of the attribute through time. We did this last step only for our main attributes, which are existed before we merged our dataset with meteorological data. Because while we were merging our dataset with the weather data we knew that, we do not have the same number of measurements within the weather data. To have the equal number of measurements for every variable, we repeated the weather data and assumed that weather measurements stays same for the repeated portion of measurements. Therefore, this last step was only meaningful for our four main variables to show their change through time.

In our dataset, we do have two real categorical variables, which are Day/Night and Month. Other than these two variables, we have Time, Date, Sunrise, Sunset, Moonrise and Moonset. The first two are the categories for time stamps, Day/Night has two categories and represent the day and night times and month has 12 categories, which represent the months. The other two variables, Time and Date are the corresponding dates and times of measurements. The last four variables are the meteorological times for sunrise, sunset, moonrise and moonset for the specific date of measurements occur. Therefore, for our categorical variables, we plotted the frequencies for every category and tried to see the distribution of the categories. We tried to measure the consistency of categories within a variable and made some comments about the results.

3.1 Plots of Categorical Variables

3.1.1 Date

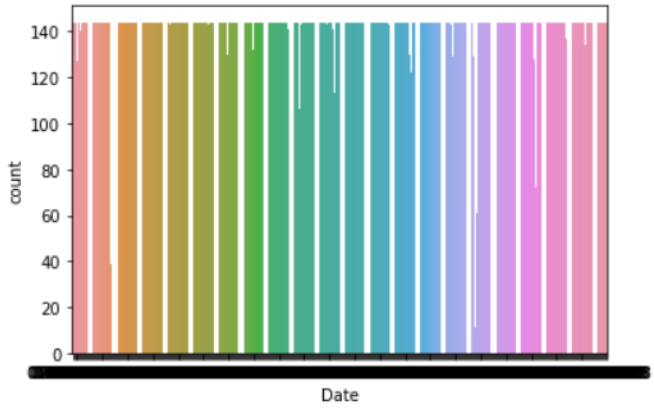


Figure 3.1.1 Frequency Plot for Date Attribute

Date variable is represent dates of the corresponding measures. Since we have measures for every 10 minutes, ideally we should have 144 measures for each day. This means our frequency for days should be equal and 144 but as we mentioned on the previous part of analysis we do not have 144 measurements for every date. Still, our plot shows that frequencies of days are very close to ideal case in general.

3.1.2 Time

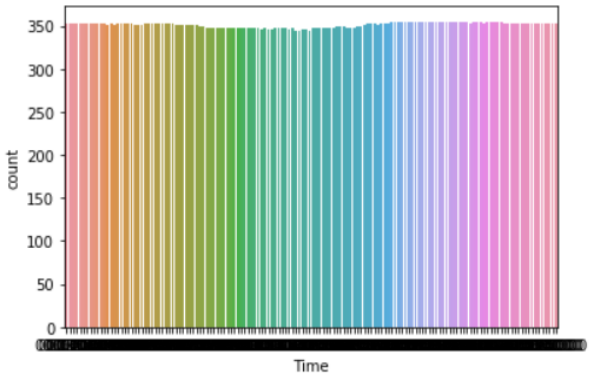


Figure 3.1.2 Frequency Plot for Time Attribute

Similar to the Date variable case, ideally we should have 365 different times because we have a 1-year length data, but as mentioned on the previous analysis, our maximum frequency for times is 355. Frequencies seem to close to each other so that plot seems very similar to a uniform distribution plot.

3.1.3 Month

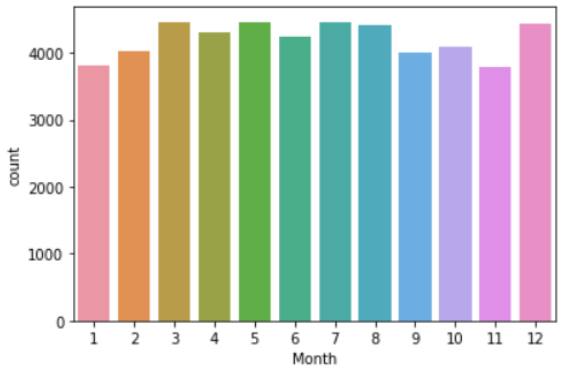


Figure 3.1.3 Frequency Plot for Month Attribute

Since every month has different number of days and due to different number of observations per day and per time intervals as mentioned above, frequencies are not same but similar.

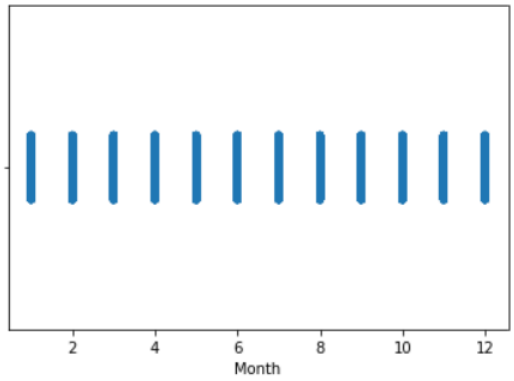


Figure 3.1.4 Strip Plot for Month Attribute

3.1.4 Day

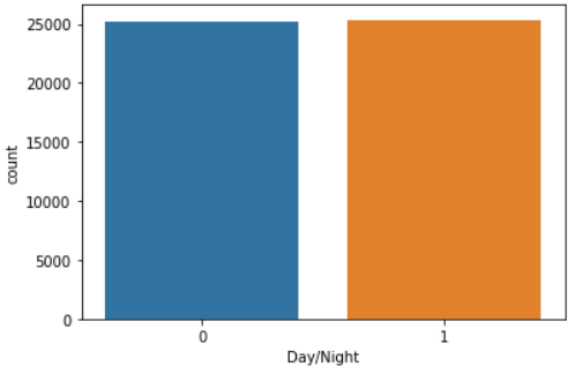


Figure 3.1.5 Frequency Plot for Day/Night Attribute

As we mentioned on the previous analysis Day and Night frequencies should also be equal in an ideal scenario but even though they are not the same they are very close to each other as our plot represents it.

3.1.5 Meteorological Variables

As we described on previous analysis these four variables are rather meteorological and we did not examine them in detail.

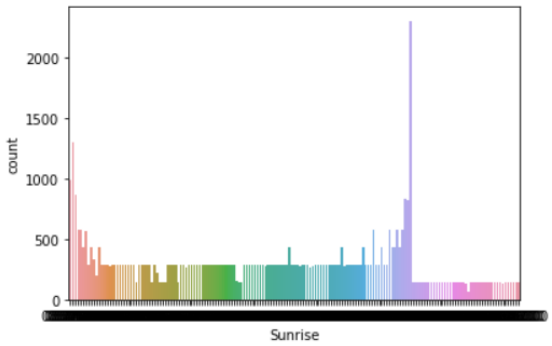


Figure 3.1.6 Frequency Plot for Sunrise Attribute

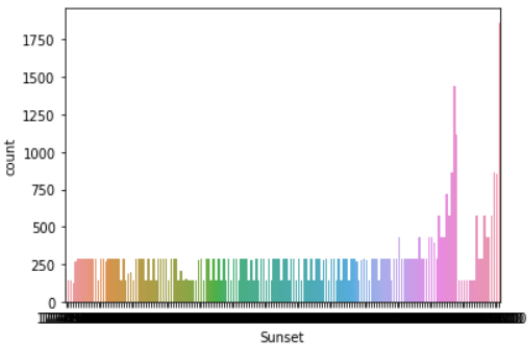


Figure 3.1.7 Frequency Plot for Sunset Attribute

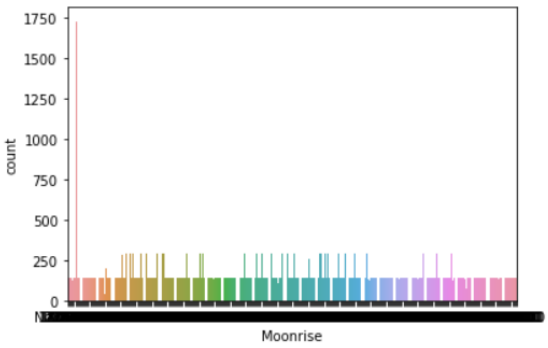


Figure 3.1.8 Frequency Plot for Moonrise Attribute

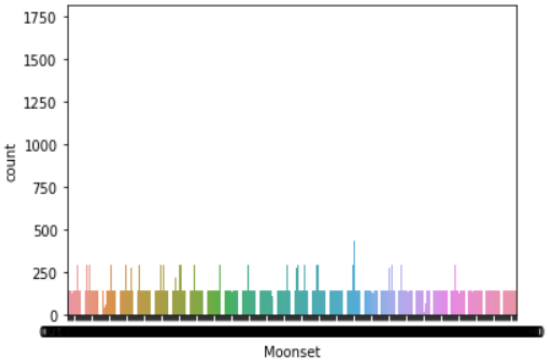


Figure 3.1.9 Frequency Plot for Moonset Attribute

3.2 Plots of Numerical Variables

3.2.1 LV Active Power (kW)

LV Active Power (kW) is our target variable. It basically represents the measured electric production of the wind turbine in the corresponding date and time.

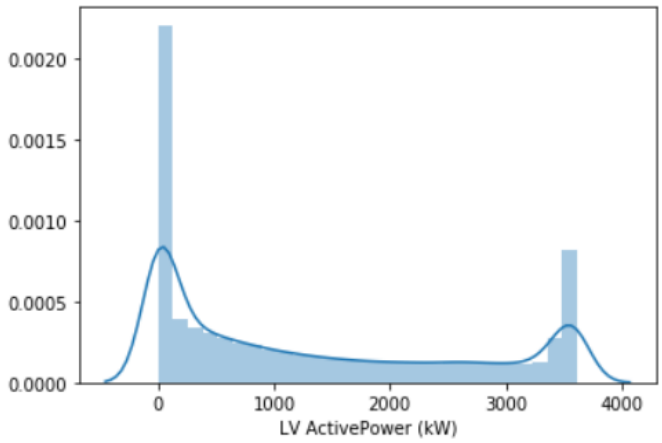


Figure 3.2.1.1 Histogram of LV Active Power (kW) Attribute

Distribution of the attribute seem to have 2 peaks. The first peak is around the 0 and it is because our dataset includes quite a lot zeros. There are different reasons for this 0 values, such as wind turbine's inability to produce electricity in certain wind speeds (low and high), probably another reasons for production inability and probably certain conditions that causes inability of measurement device. This first peak means we have 0 as most frequent value in our dataset. Despite the first peak as electric production increases the frequencies seem to decrease except around the second peak.

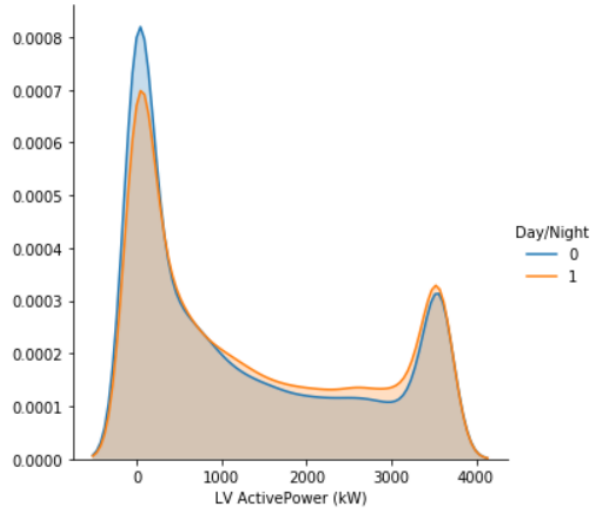


Figure 3.2.1.2 Distribution of LV Active Power (kW) Attribute as Day and Night

Electric production does not differ in day and night times. Only difference is that night measurements have more 0 values.

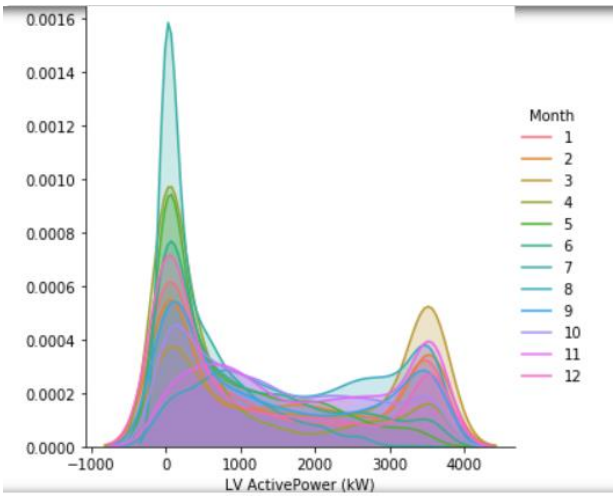


Figure 3.2.1.3 Distribution of LV Active Power (kW) Attribute in Months

Electric production also distributes similarly in different months except for 9th and 11th months do not have a peak for 0 values and 7th month seem to have highest peak in other words most number of 0 values.

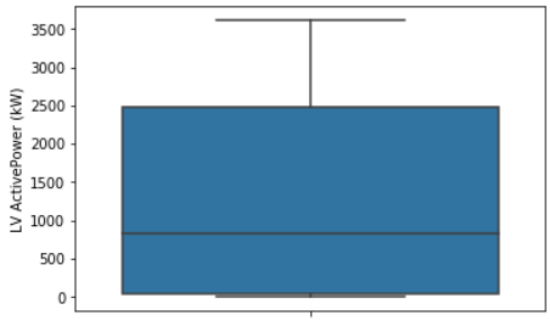


Figure 3.2.1.4 Boxplot of LV Active Power (kW) Attribute

Our boxplot shows that the data is distributed generally between 25%-75% range. There seem to be no outliers . Our median or 50% value -825- and mean value -1307- seems highly different. For generality purposes we can rely much on median in this case since mean is highly effected by large values.

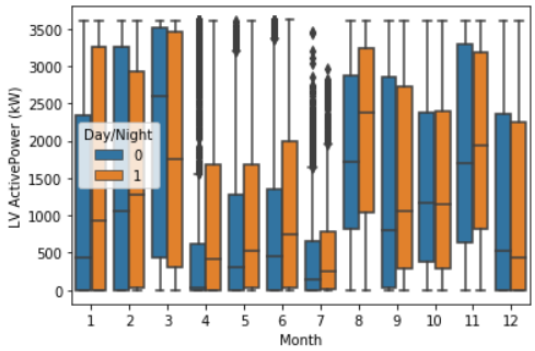


Figure 3.2.1.5 Boxplot of LV Active Power (kW) Attribute for Subdivisions

We can see here that throughout the months, day and night productions are different within the same month. Generally, Day productions (1) seem to have higher variance than the night

productions. We can see that in months 5 and 6 additional to the month 4-7 ,have lots of outliers but only visible in night productions except for month 7.

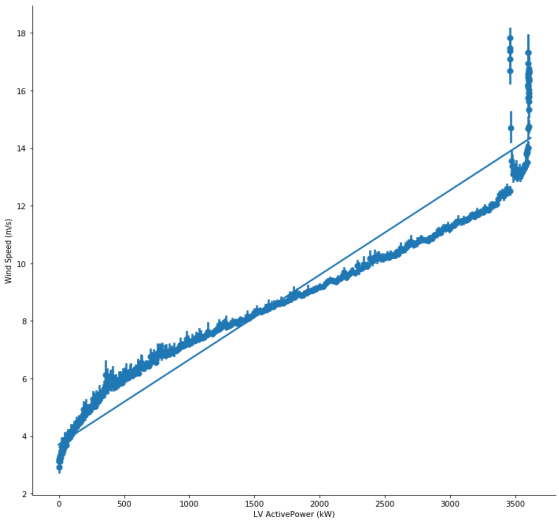


Figure 3.2.1.6 Linear model Plot of LV Active Power (kW) – Wind Speed

As we expected before our analysis, there is a positive linear relationship between electric production and wind speed variables.

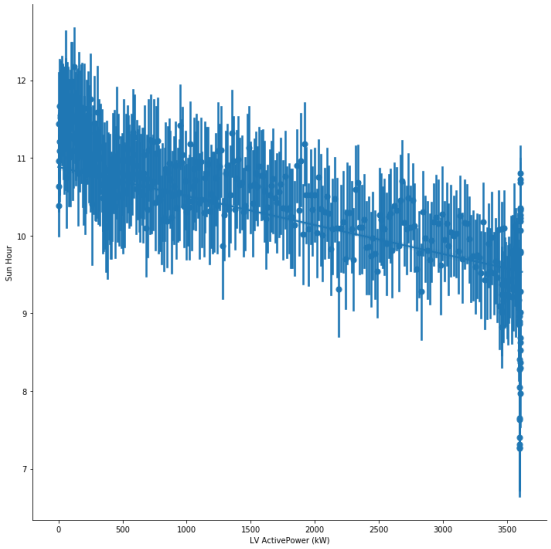


Figure 3.2.1.7 Linear model Plot of LV Active Power (kW) – Sun Hour

An unexpected result, there seem to be a negative linear relationship between electric production and Sun Hour variable

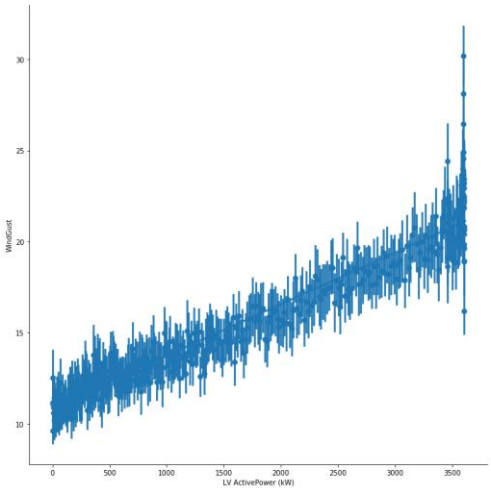


Figure 3.2.1.8 Linear model Plot of LV Active Power (kW) – Wind Gust

Electric production and Wind Gust variable have a positive linear relationship. Wind gust represents sudden increases in the wind speed.

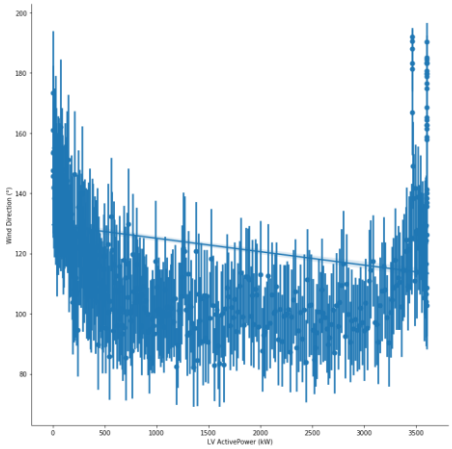


Figure 3.2.1.9 Linear model Plot of LV Active Power (kW) – Wind Direction

There seem to be a nonlinear relationship, rather a quadratic one between Electric production and Wind Direction but, wind direction values are angel values so to interpret this plot as it is will not be appropriate.

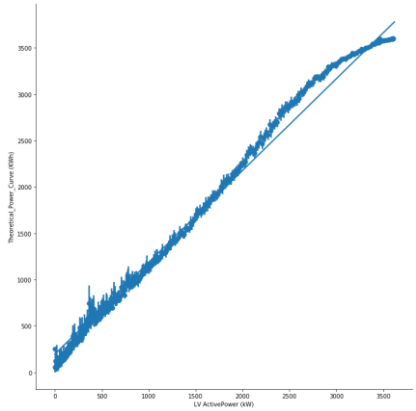


Figure 3.2.1.10 Linear model Plot of LV Active Power (kW) - Theoretical_Power_Curve (KWh)

As expected there is a very strong linear relationship between electric production values of the wind turbine and theoretical calculations of electric production of the wind turbine.

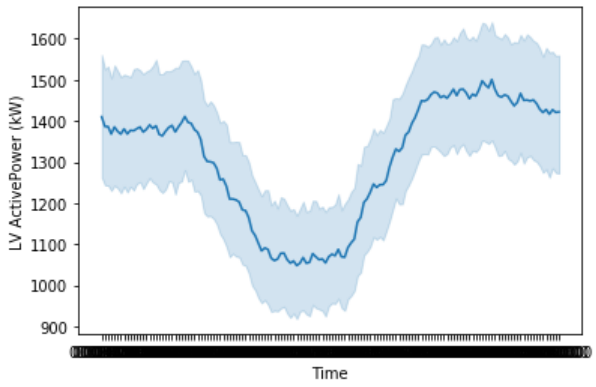


Figure 3.2.1.11 Time Series of LV Active Power (kW)

The average electric production for every date. It is the average of the 144 measurements for every day. The shaded area represent the diversity of the values around the mean. The midyear seems to have less production than the other times of the year.

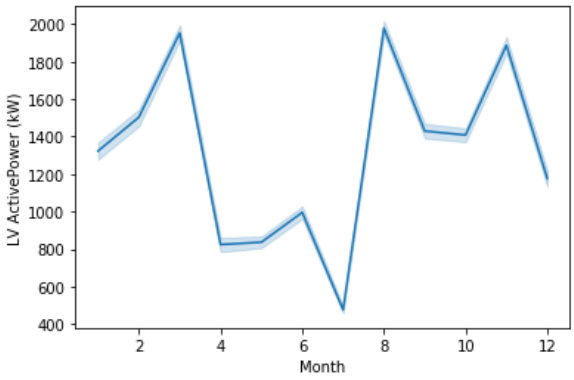


Figure 3.2.1.12 Time Series of LV Active Power (kW) in Months

3.2.2 Wind Speed (m/s)

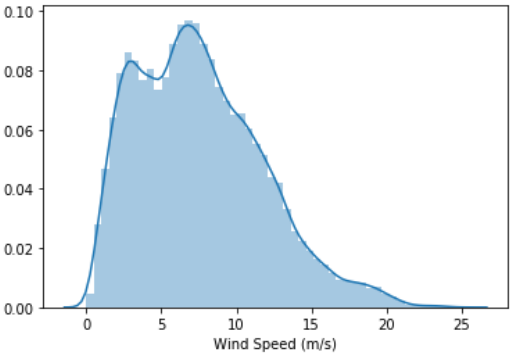


Figure 3.2.2.1 Histogram of Wind Speed (m/s)

Distribution of the wind speed has 2 peaks.

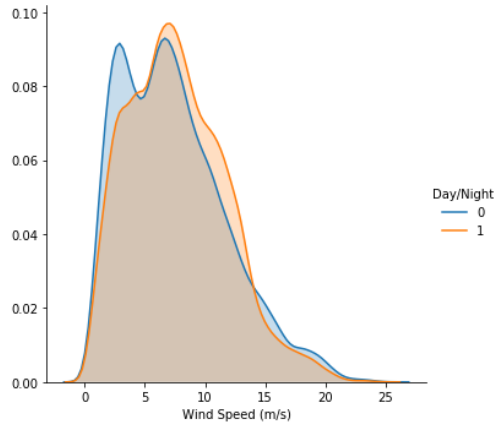


Figure 3.2.2.2 Distribution of Wind Speed (m/s) with Day/Night

When looking in to more detail we can see that our original distribution of the values has 2 peaks but the Day sub category has only one peak. The first peak on the general plot seem to be caused by the Night sub category.

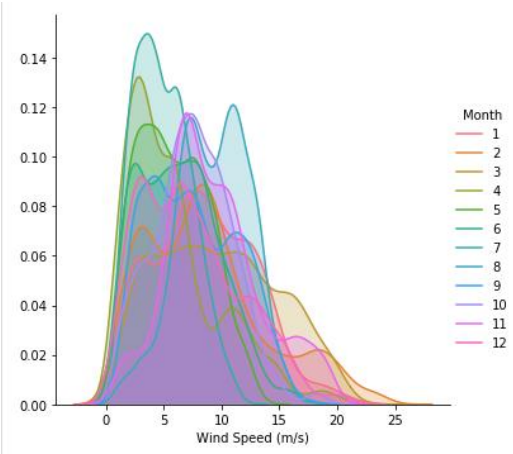


Figure 3.2.2.3 Distribution of Wind Speed (m/s) with Months

When looking into months, it is far more interesting. We can see there are very different distributions in monthly evaluations

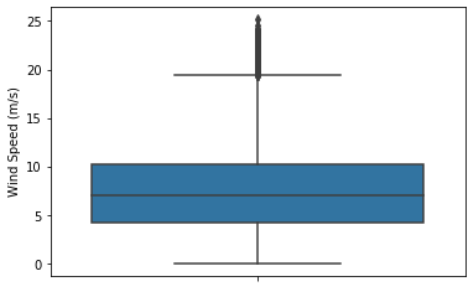


Figure 3.2.2.4 Boxplot of Wind Speed (m/s)

We can see that wind speed data has lots of outliers. After checking statistics, we can see median and mean are similar in this data.

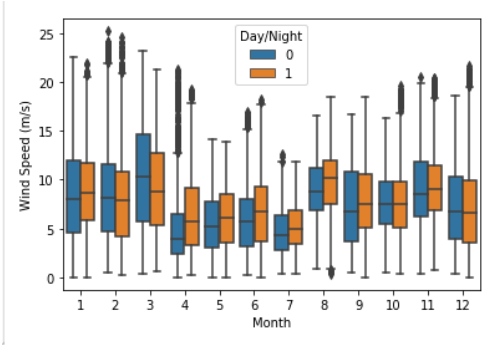


Figure 3.2.2.5 Boxplot of Wind Speed (m/s) with subdivision

In month from 4 to 7 it seems that wind speeds decrease. It may be because of the seasons effect.

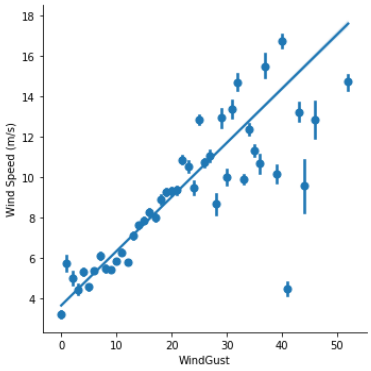


Figure 3.2.2.6 Linear model plot of Wind Speed (m/s)-Wind Gust

Wind Gust and Wind Speed seem to have positive linear relationship.

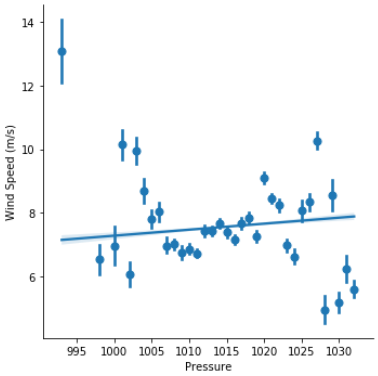


Figure 3.2.2.7 Linear model plot of Wind Speed (m/s)-Pressure

Pressure and wind speed does not seem to have a weak linear relationship.

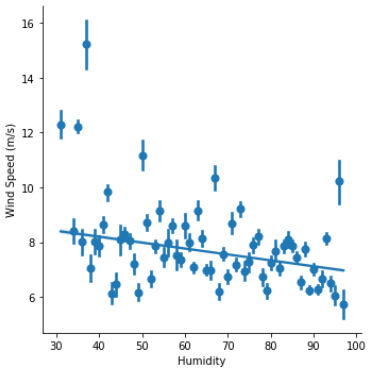


Figure 3.2.2.8 Linear model plot of Wind Speed (m/s)-Density

Humidity and Wind Speed seem to have a negative linear relationship.

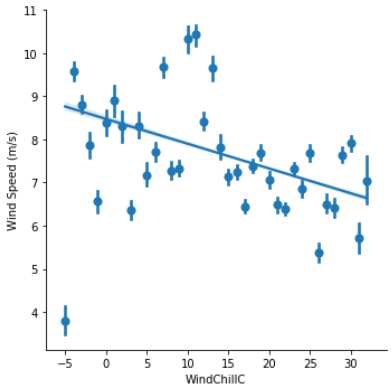


Figure 3.2.2.9 Linear model plot of Wind Speed (m/s)-WindChillC

WindChill and Wind Speed seem to have a negative linear relationship.

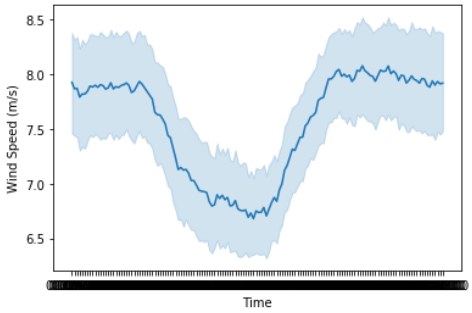


Figure 3.2.2.10 Time Series plot of the Wind Speed

It shows a very similar distribution to LV Active Power -electric production- data.

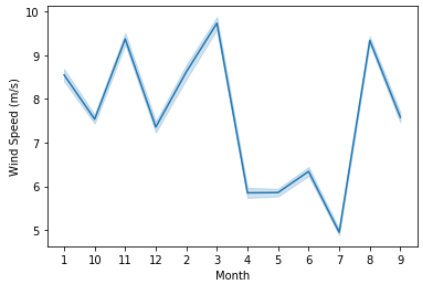


Figure 3.2.2.11 Time Series plot of the Wind Speed in Months

3.2.3 Theoretical Power Curve (KWh)

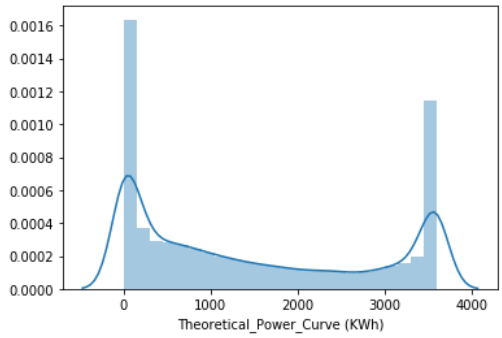


Figure 3.2.3.1 Histogram of Theoretical Power Curve Attribute

Since this variable is a theoretical prediction of LV Active Power variable most of the plots are very similar. We found differences in linear relationship with other variables.

Distribution of Theoretical Power Curve is very similar to LV Active Power as expected.

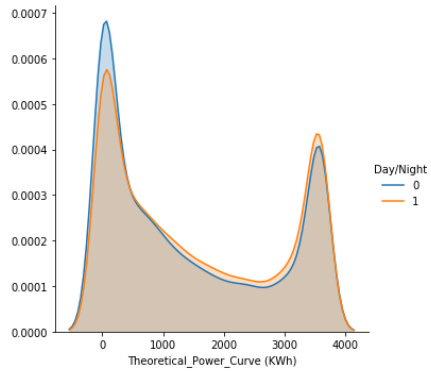


Figure 3.2.3.2 Distribution of Theoretical Power Curve Attribute as Day/Night

As in LV Active Power attribute, Day and Night distributions are similar.

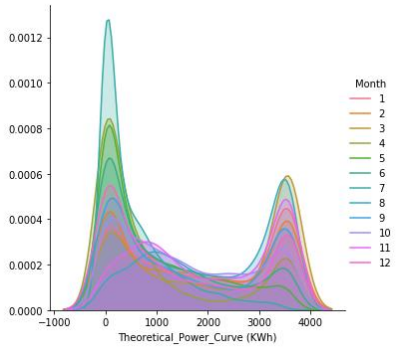


Figure 3.2.3.3 Distribution of Theoretical Power Curve Attribute with Months

Distribution in months are also very similar to LV Active Power variable's distribution.

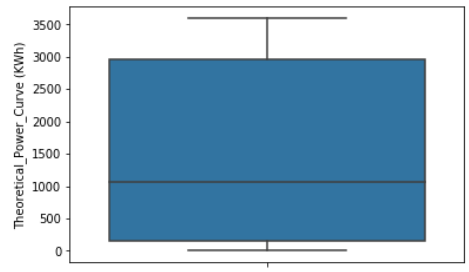


Figure 3.2.3.4 Boxplot of Theoretical Power Curve Attribute

Boxplots and its variations are also very similar to LV Active Power attribute's plots

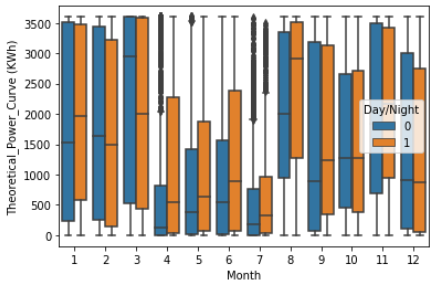


Figure 3.2.3.5 Boxplot of Theoretical Power Curve Attribute in Subdivisions

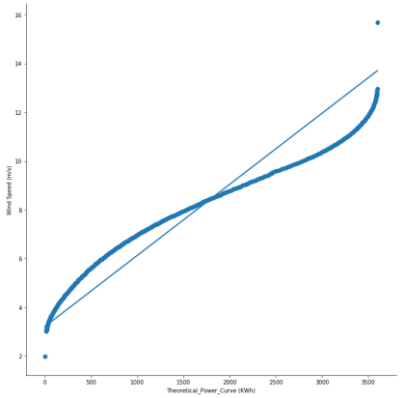


Figure 3.2.3.6 Linear model plot of Theoretical Power Curve – Wind Speed

It seems a bit more diverse than the LV Active Power-Wind Graph, it may be because of the theoretical formula does not imply enough weight to the wind but in real life the wind speed actually matters more.

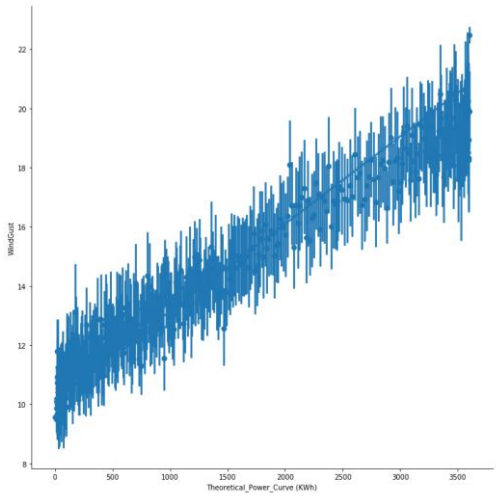


Figure 3.2.3.6 Linear model plot of Theoretical Power Curve – Wind Gust

Wind gust and theoretical value seem to have strong linear relationship. The relationship is stronger than the real production values' relationship with wind gust but also it involves more error than the real production values-Wind gust plot.

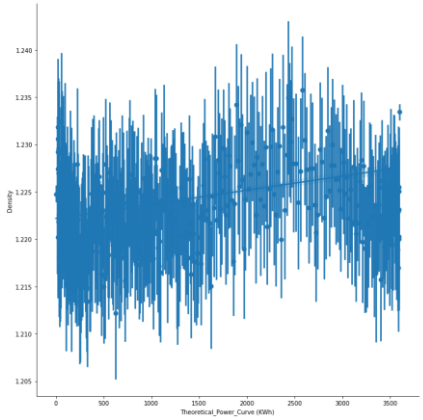


Figure 3.2.3.7 Linear model plot of Theoretical Power Curve – Density

Density and Theoretical value seem to have linear relationship. The real production values little or no relationship with the Density. These differences exist theoretical formula directly related with the air density.

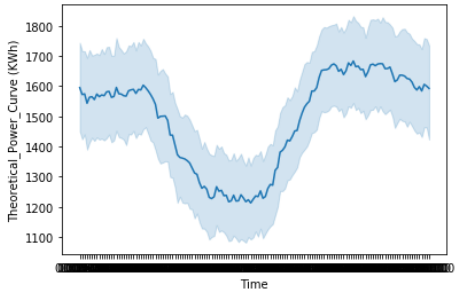


Figure 3.2.3.8 Time Series Plot of Theoretical Power Curve

This plot is also very similar to the LV Active Power's Time Series plot.

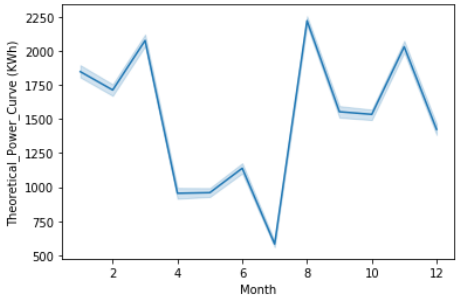


Figure 3.2.3.9 Time Series Plot of Theoretical Power Curve

3.2.4 Wind Direction (°)

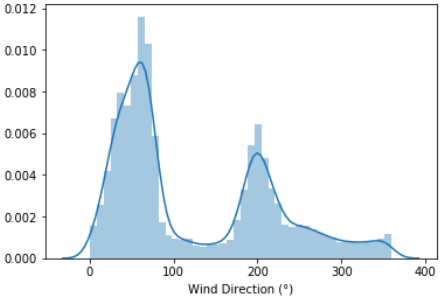


Figure 3.2.4.1 Histogram of Wind Direction Attribute

Our wind direction is angles in terms of degrees. So it has to be between 0-360. There are two peaks, one is around 100 degrees and the other peak is very close to 200 degrees.

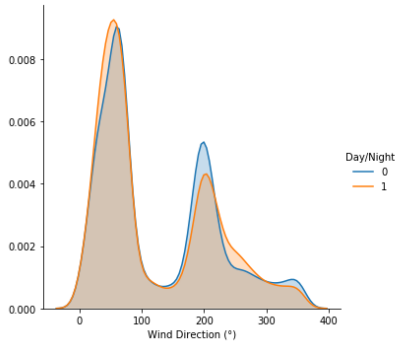


Figure 3.2.4.2 Distribution of Wind Direction Attribute with Day/Night

Day and night distributions are very similar.

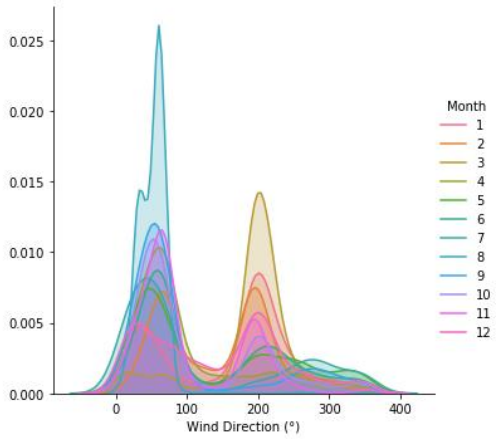


Figure 3.2.4.3 Distribution of Wind Direction Attribute within Months

Other than months 4 and 9 other seem to distribute similarly, both month 4 and 9 have really high peaks, they may be the reason of the peaks at the general distribution.

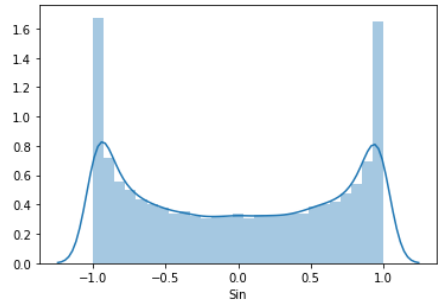


Figure 3.2.4.4 Histogram of Sinus of Wind Direction Attribute

Since in degree terms the degrees 0 and 360 means the same, to change it and have a normalization effect. We calculated the sin(x) for our wind direction angles. To have values between -1 and 1.The distribution seems a little bit different but it also has 2 peaks and both of them around the extremes. The reason of it probably our angle distribution peaks was around 100 and 200 and sin90 and sin180 are the extreme values. Probably we have a high frequency of 90 and 180 degrees.

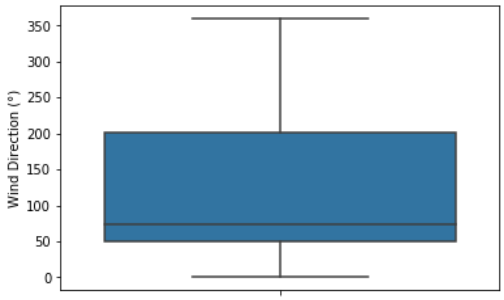


Figure 3.2.4.5 Boxplot of Wind Direction Attribute

Probably as we seen on the distribution plots, because of the high frequency, of the values around 100 and below our median is 73 but our mean is 123. We have a relatively high standard deviation.

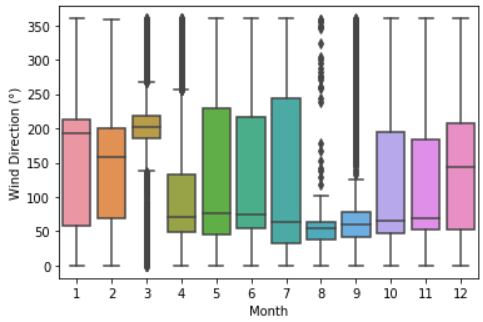


Figure 3.2.4.6 Boxplot of Wind Direction Attribute within Subdivisions

Day and night seem to have very similar distribution Months 3-4 and 8-9 has many outliers.

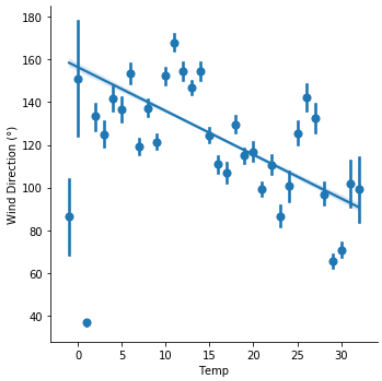


Figure 3.2.4.7 Linear model plot of Wind Direction-Temperature Attributes

These graph of wind direction with other variables cannot interpreted as same as other variables because angles are not like numeric values exactly. 160 degree does not imply it is bigger than 120 degree It actually implies the directions.

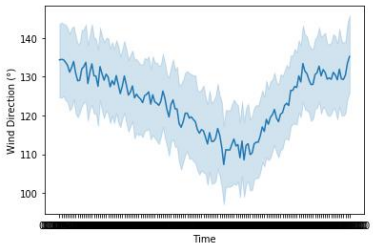


Figure 3.2.4.8 Time Series plot of Wind Direction

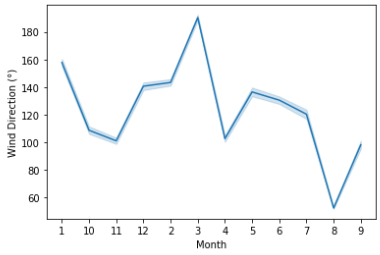


Figure 3.2.4.9 Time Series plot of Wind Direction in Months

3.2.5 Temp

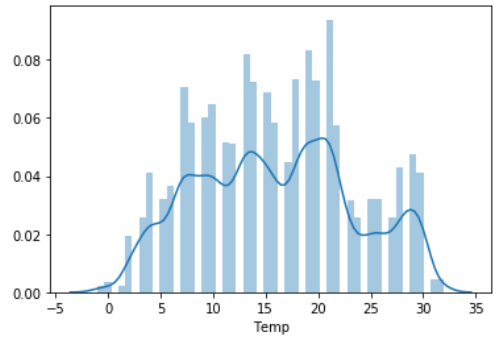


Figure 3.2.5.1 Histogram of Temp Attribute

Temp basically is the temperature of the place which measurement take place in corresponding time and date.

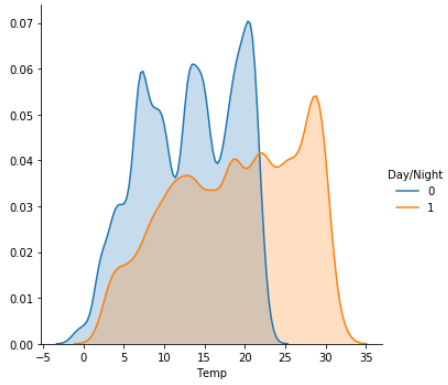


Figure 3.2.5.2 Distribution of Temp Attribute for Day/Night Features

Day and night distributions seem very different as expected.

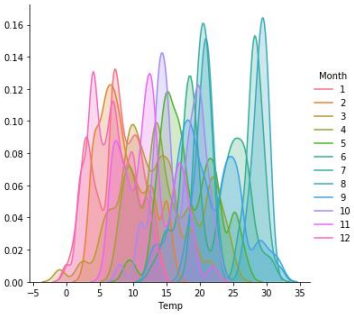


Figure 3.2.5.3 Distribution of Temp Attribute in Months

Different moths shows different distributions as expected.

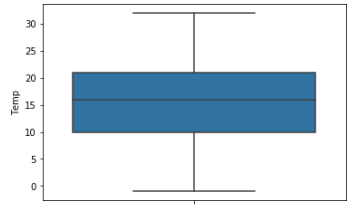


Figure 3.2.5.4 Boxplot of Temp Attribute

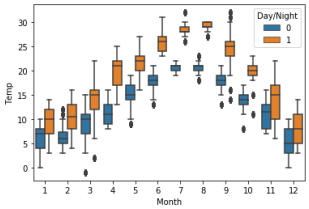


Figure 3.2.5.5 Boxplot of Temp Attribute with Subdivisions

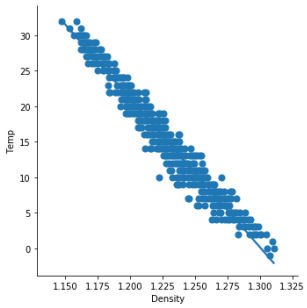


Figure 3.2.5.6 Linear Model Plot Temp-Density

Temp and Air Density seem to have a strong negative linear relationship.

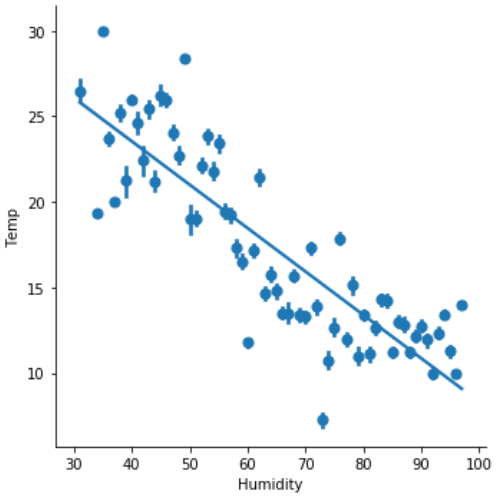


Figure 3.2.5.7 Linear Model Plot Temp-Humidity

Temperature and humidity seem to have a negative linear relationship.

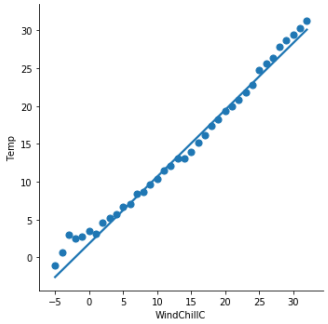


Figure 3.2.5.8 Linear Model Plot Temp-WindChill

Temperature and windchill has a very strong linear relationship.

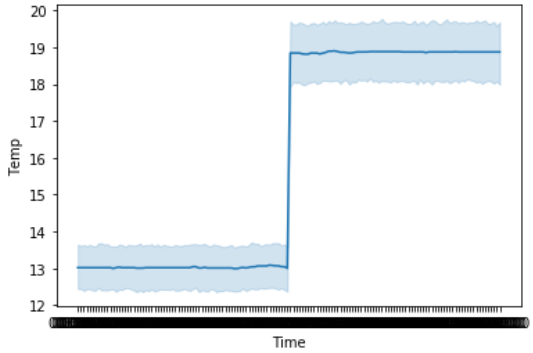


Figure 3.2.5.9 Time Series Plot of Temp Attribute

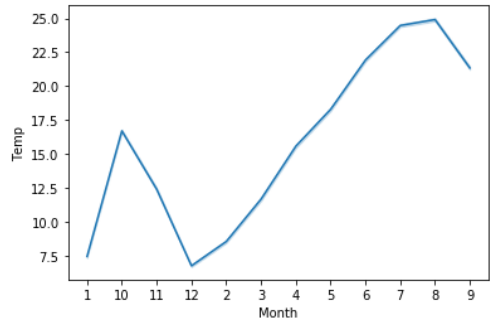


Figure 3.2.5.10 Time Series Plot of Temp Attribute

3.2.6 SunHour

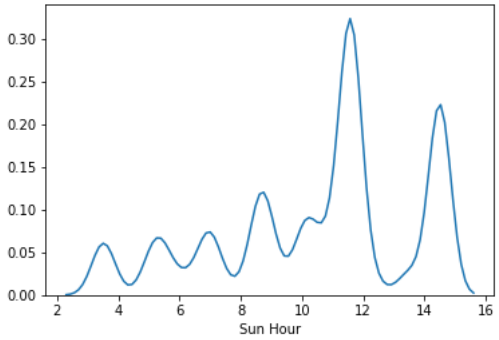


Figure 3.2.6.1 Distribution of SunHour Attribute

In this Graph, we see that distribution of SunHour. Between 10-12 and 14-16 is peak values. In this time sunhour values density is so high.

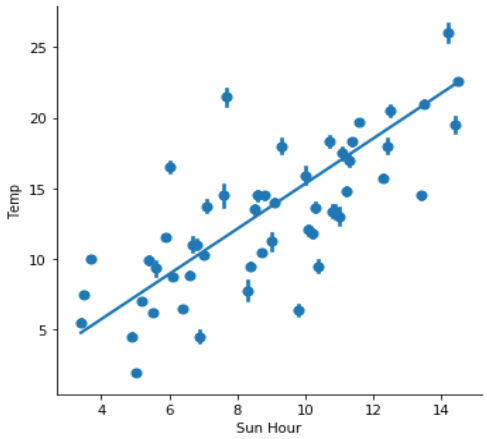


Figure 3.2.6.2 Linear Model Plot SunHour- Temperature

In this graph, we see that Sun Hour and Temperature have a linear relationship.

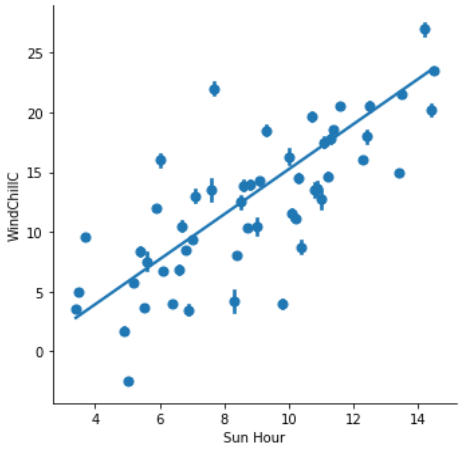


Figure 3.2.6.3 Linear Model Plot SunHour- WindChillC

In this graph, we see that Sun Hour and WindChillC have a strong linear relationship.

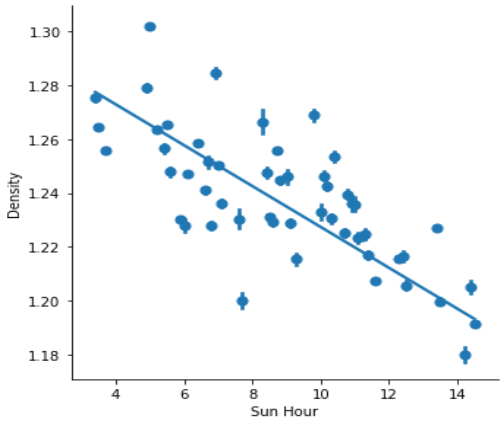


Figure 3.2.6.4 Linear Model Plot SunHour- Density

In this graph, we see that Sun Hour and Density have a strong negative linear relationship.

3.2.7 Moon Illumination

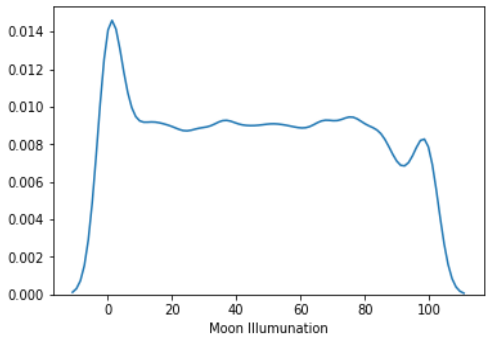


Figure 3.2.7.1 Histogram of Moon Illumination Attribute

In this Graph, we see that distribution of MoonIllumination. After the density begins to form, there is a rapid increase to the peak. Later, although a certain amount of decrease appears, the density is generally high.After MI is 100, the intensity reaches zero

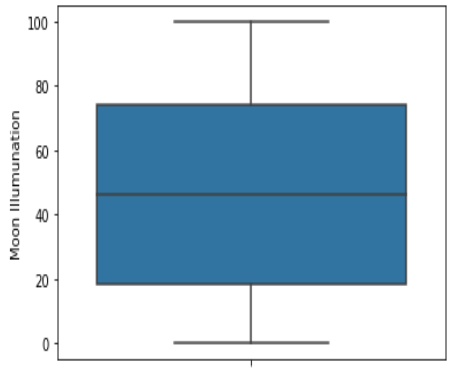


Figure 3.2.7.2 Boxplot of Moon Illumination Attribute

In this Graph, we see that Our boxplot shows that data is distributed generally between 20%-80% range. It has a balanced distribution. There seem to be no outliers in the boxplot.

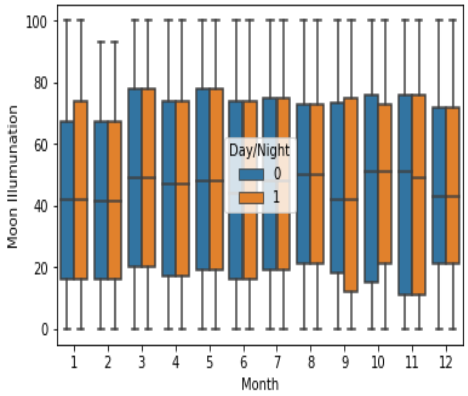


Figure 3.2.7.3 Boxplot of Moon Illumination Attribute Day/Night all month

In this graph, we see that We can also see here that Moon Illumunation values the months, day and night. In generally, Moon Illumunation value generally has a homogeneous distribution.

We did not find any meaningful linear relationship between moon illumination variable and other variables. Still to make an example we put some linear model plots.

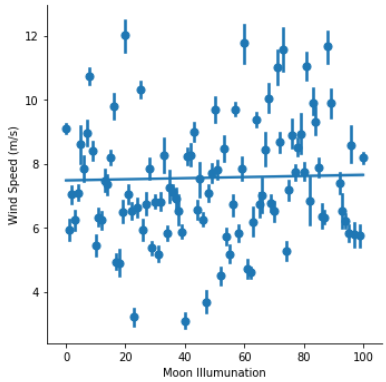


Figure 3.2.7.4 Linear Model Plot of Moon Illuminaton and Wind Speed

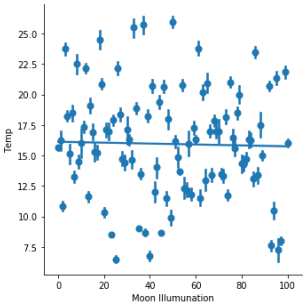


Figure 3.2.7.4 Linear Model Plot of Moon Illuminaton and Temp

3.2.8 DewPoint

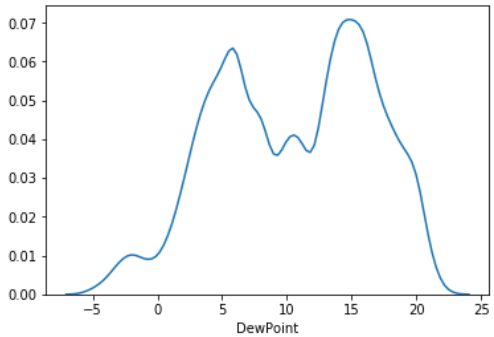


Figure 3.2.8.1 Histogram of Dew Point Attribute

In this Graph, we see that distribution of DewPoint. When DewPoint is -5 , density starts to form but at low values. At 5 and 15 values is peak values. Density reach so high levels. Then , density start to decrease and it is 0 at the DewPoint is 25.

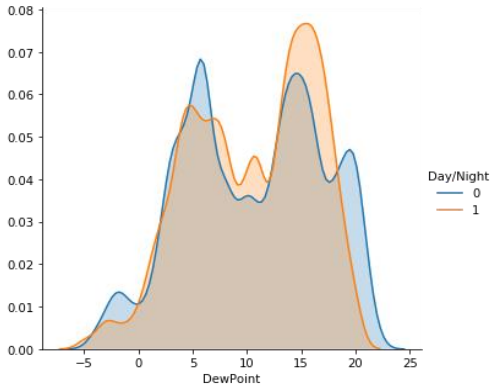


Figure 3.2.8.2 Histogram of Dew Point Attribute Day/Night

In this graph, we see that distribution of Dewpoint as DAY and NIGHT. They have approximately the same distribution. There is not much change in night and day.

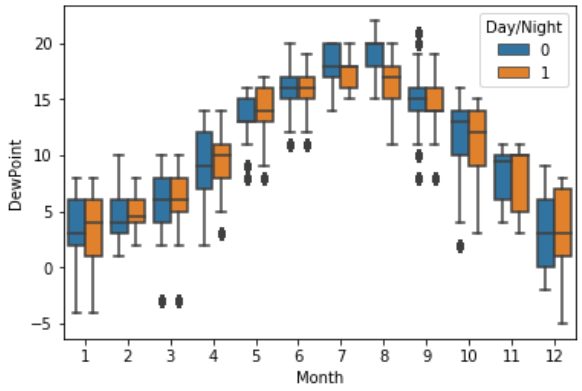


Figure 3.2.8.3 Boxplot of DewPoint Attribute with Subdivisions

In this Graph, we see that Night and Day distribution of DewPoint value monthly. After the 4th month, DewPoint values increase and peak value is 7th month . Then , Values start to decrease. Night and day values are generally in the same range. But in some months it shows differences like 7 and 8th month.

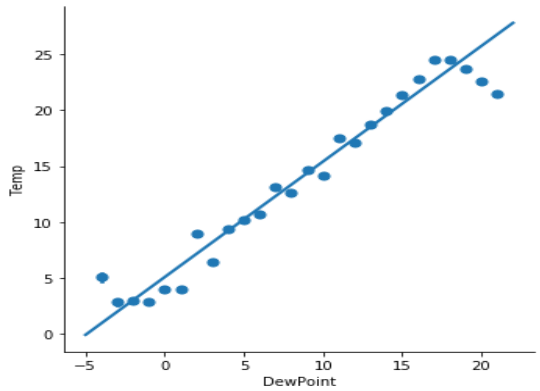


Figure 3.2.8.4 Linear Model Plot DewPoint-Temp

In this graph, we see that DewPoint and Temperature have a strong linear relationship. It means that they directly connected each other.

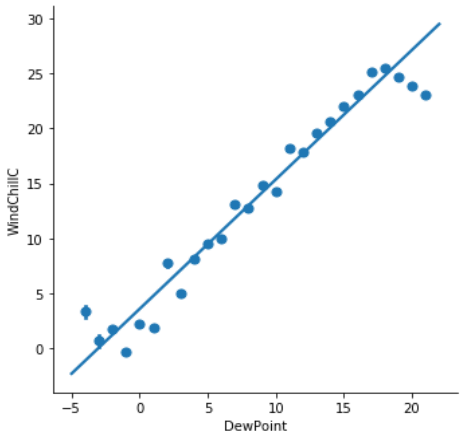


Figure 3.2.8.5 Linear Model Plot DewPoint-WindChillC

In this graph, we see that DewPoint and WindChill have a strong linear relationship. It means that they directly connected each other.

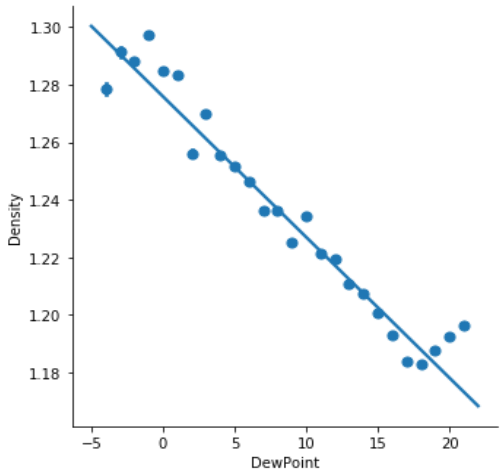


Figure 3.2.8.6 Linear Model Plot DewPoint-Density

In this graph, we see that DewPoint and Density have a strong negative linear relationship. It means that they directly connected each other.

3.2.9 WindChillC

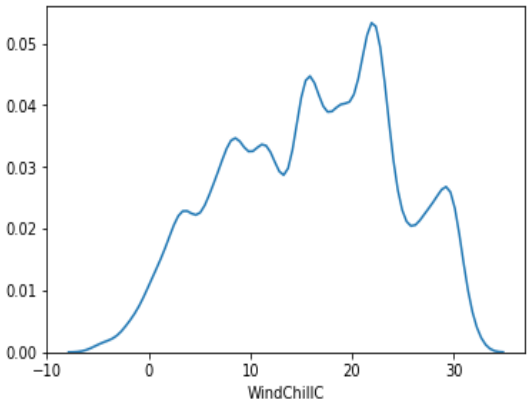


Figure 3.2.9.1 Histogram of Wind ChillC Attribute

In this Graph, we see that distribution of WindChillC. When WindChillC is -10, density starts to form but at low values. At 20 is peak values. Density reach so high levels. Then , density start to decrease and it is 0 at the WindChillC is 35.Generally , WindChillC values have high density ratio.

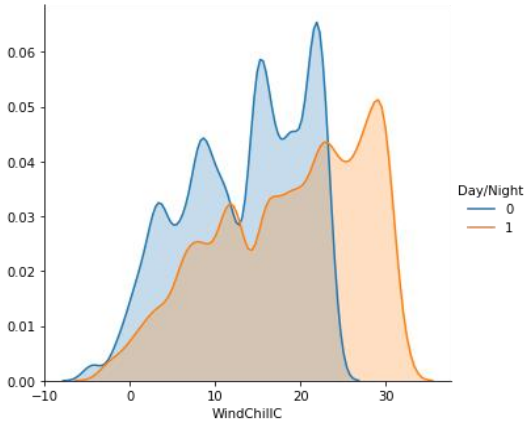


Figure 3.2.9.2 Histogram of Wind ChillC Attribute Day/Night

In this graph, we see that distribution of WindChillC as DAY and NIGHT. At night, windchill values can reach 35, but this value does not exceed 25 throughout the day.

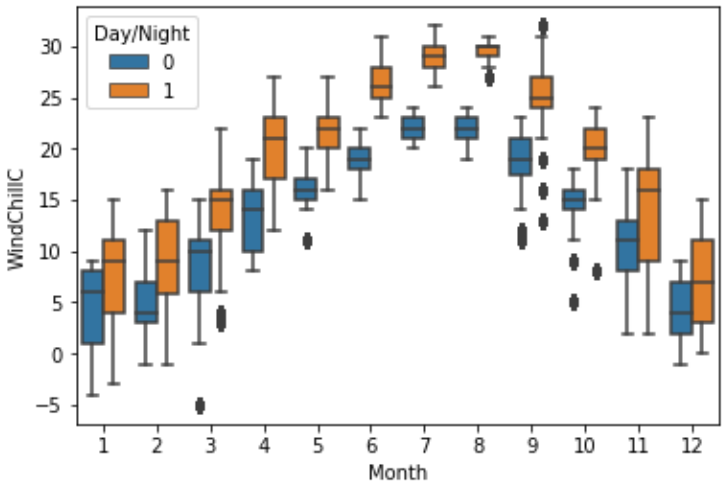


Figure 3.2.9.3 Boxplot of Wind ChillC Attribute with Subdivisions

In this Graph, we see that Night and Day distribution of WindChill value monthly. After the 4th month, WindChill values increase and peak value is 7th month . Then , Values start to decrease. Night and Day values generally have different ranges. The range of values becomes narrower as you approach the peak of the chart. There are so much outlier value.

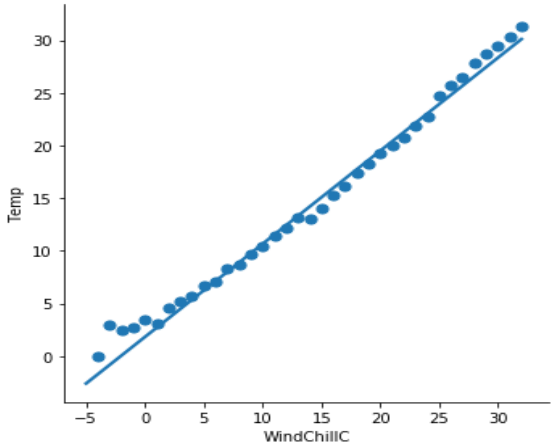


Figure 3.2.9.4 Linear Model Plot WindChillC-Temperature

In this graph, we see that WindChill and Temperature have a strong linear relationship. It means that they directly connected each other.

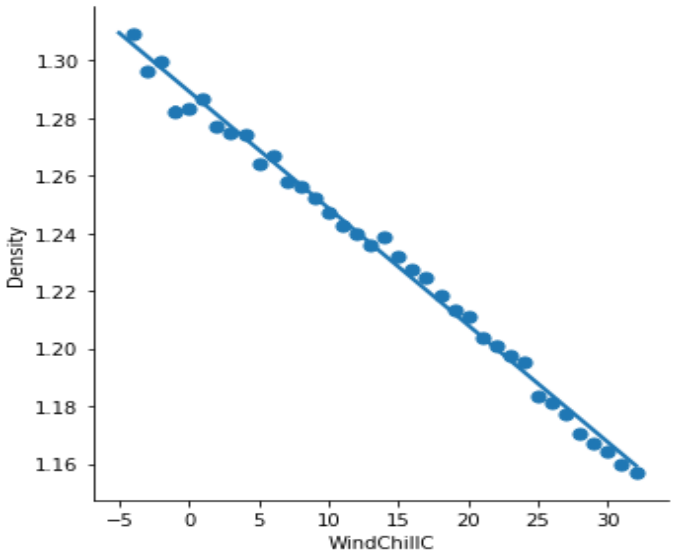


Figure 3.2.9.5 Linear Model Plot WindChillC-Density

In this graph, we see that WindChill and Density have a strong negative linear relationship. It means that they directly connected each other.

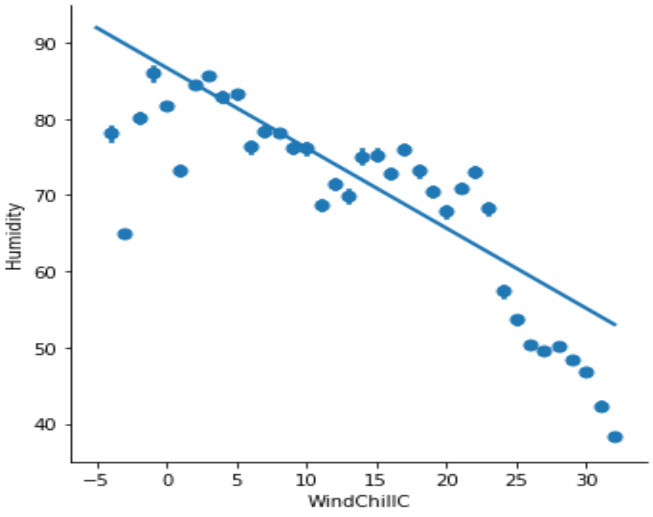


Figure 3.2.9.6 Linear Model Plot WindChillC-Humidity

In this graph, we see that WindChill and Humidity have a negative linear relationship.

3.2.10 WindGust

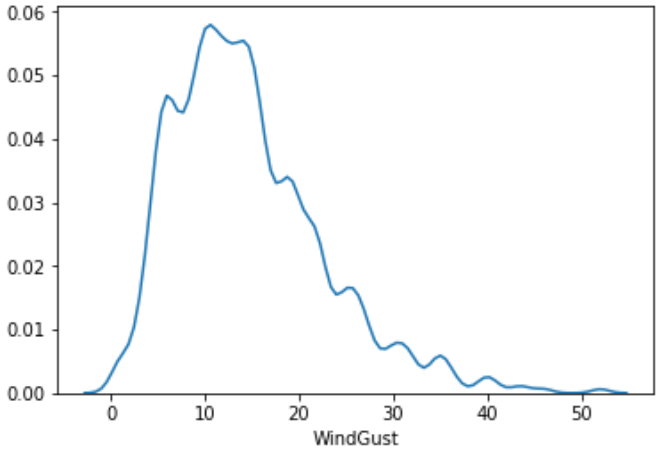


Figure 3.2.10.1 Histogram of WindGust Attribute

In this Graph, we see that distribution of WindGust. When WindGust is -5, density starts to form but itincreasing rapidly. At 15 is peak values. Density reach so high levels. Then , density start to decrease and it is 0 at the WindGust is 55. After reaching the peak value, the density decreases slowly

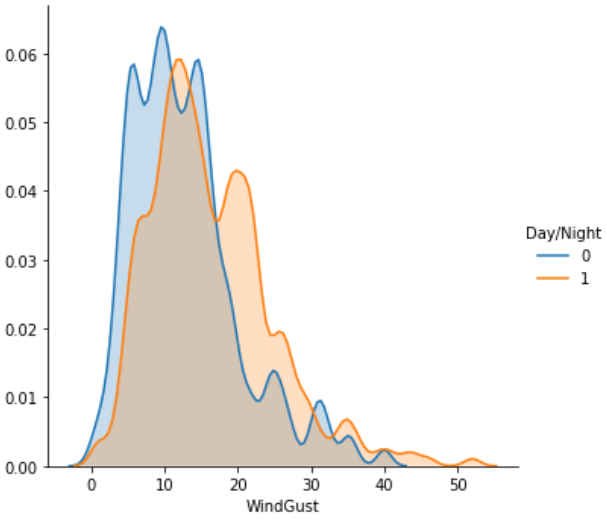


Figure 3.2.10.2 Histogram of WindGust Attribute Day/Night

In this graph, we see that distribution of WindGust as DAY and NIGHT. When WindGust values are in the 20-30 range, it has a more intense value at night. When WindGust values are in the 0-10 range, it has a more intense value at day

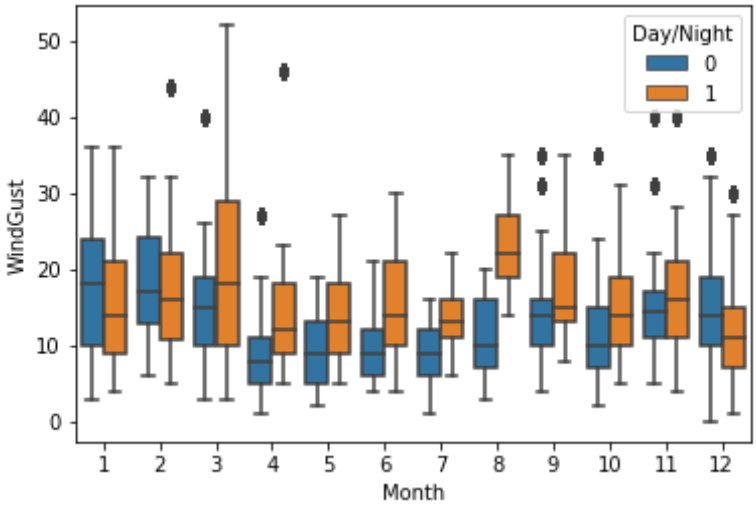


Figure 3.2.10.3 Boxplot of WindGust Attribute with Subdivisions

In this Graph, we see that Night and Day distribution of WindGust value monthly. Generally they have normal distribution. In some months, the value ranges are wider like 1,2,3. In some months, day and night, min and max values are very different like 7,8,9

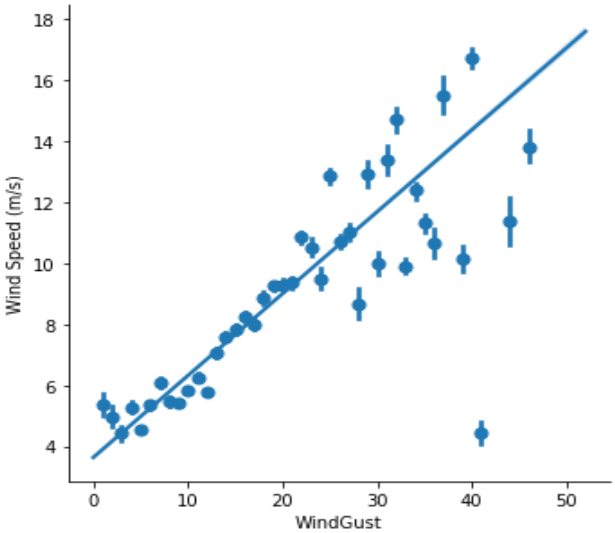


Figure 3.2.10.4 Linear Model Plot WindGust- WindSpeed

In this graph, we see that WindGust and WindSpeed have a linear relationship.

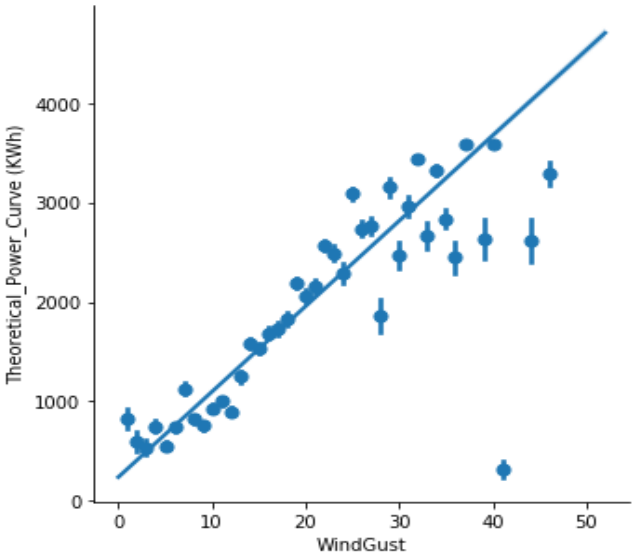


Figure 3.2.10.5 Linear Model Plot WindGust- Theoretical_Power_Curve

In this graph, we see that WindGust and Theoretical_Power_Curve (Kwh) have a strong linear relationship. It means that they directly connected each other.

3.2.11 Humidity

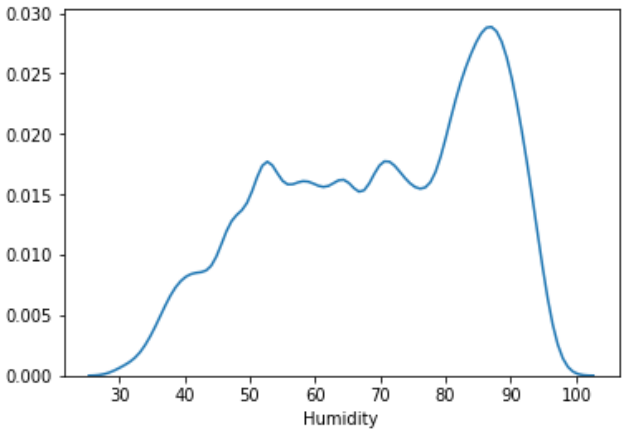


Figure 3.2.11.1 Histogram of Humidity Attribute

In this Graph, we see that distribution of Humidity. When Humidity is 30 , density starts to form . At 85 is peak value. Density reach so high levels. Then , density start to decrease fastly and it is 0 at the Humidity is 100.

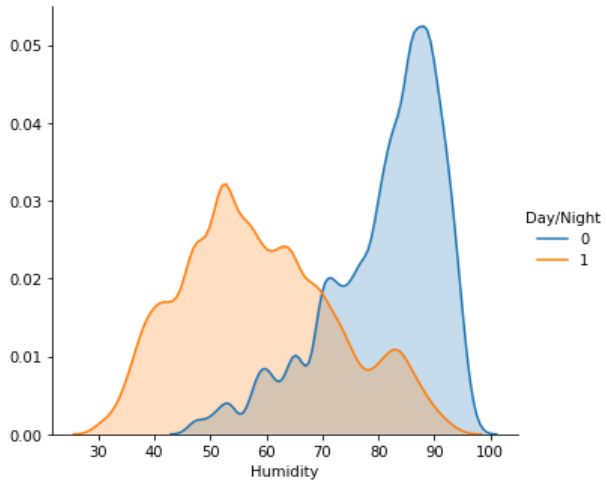


Figure 3.2.11.2 Histogram of Humidity Attribute Day/Night

In this graph, we see that distribution of Humidity as DAY and NIGHT. Humidity graph at night has a more symmetrical structure. Peak value is 55 for night. .But at day, more intensity on the right side. Peak value is 90 for day.

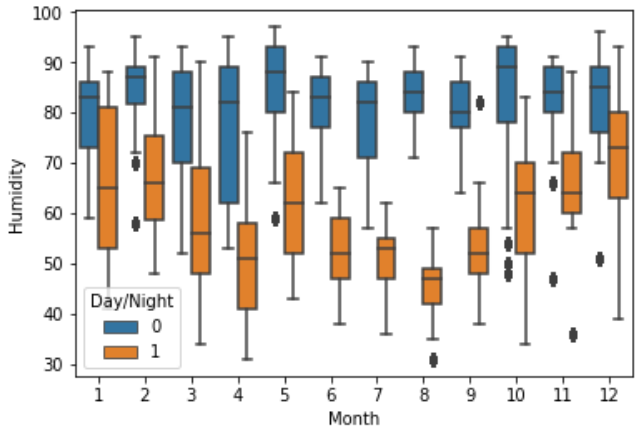


Figure 3.2.11.3 Boxplot of Humidity Attribute with Subdivisions

In this Graph, we see that Night and Day distribution of Humidity value monthly. Night and day values are generally so different range. There are so outlier values.

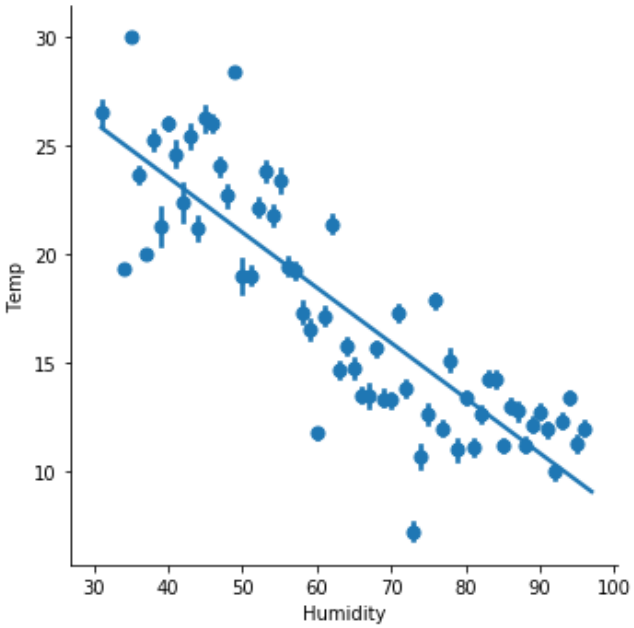


Figure 3.2.11.4 Linear Model Plot Humidity- Temperature

In this graph, we see that Humidity and Temperature have a strong negative linear relationship.

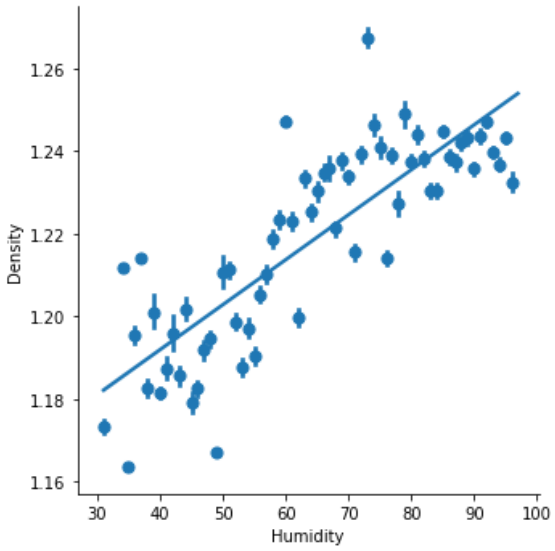


Figure 3.2.11.5 Linear Model Plot Humidity- Density

In this graph, we see that Humidity and Density have a strong linear relationship.

3.12 Pressure

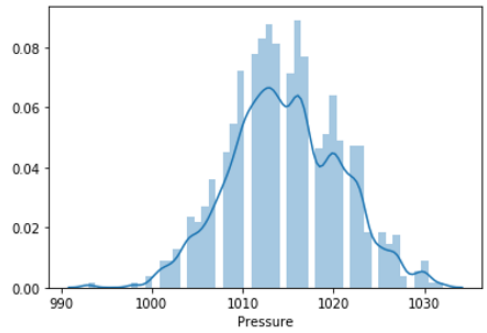


Figure 3.2.12.1 Histogram of Pressure Attribute

It seems similar to a normal distribution.

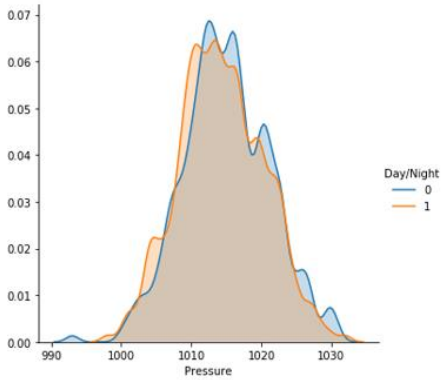


Figure 3.2.12.2 Distribution of Pressure Attribute as Day/Night
Day and night distributions are slightly different than each other.

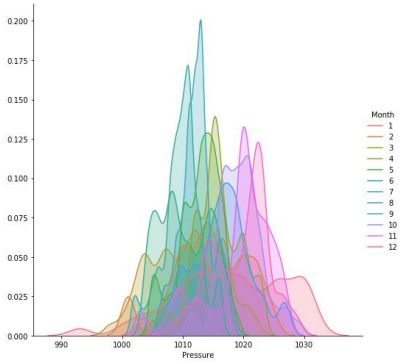


Figure 3.2.12.3 Distribution of Pressure Attribute in Months

Monthly distributions seems very different.

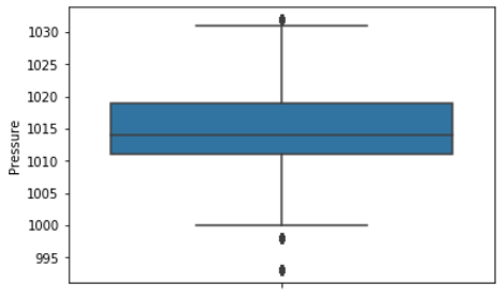


Figure 3.2.12.4 Boxplot of Pressure Attribute

It seems a rather equal distribution with a few numbers of outliers.

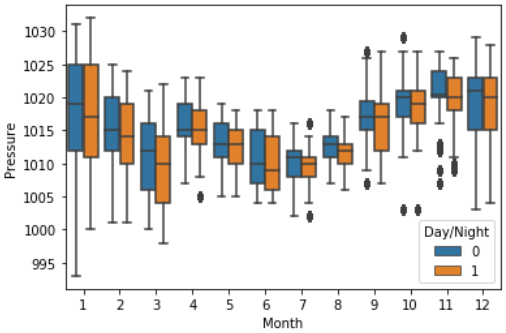


Figure 3.2.12.5 Boxplot of Pressure Attribute in Subdivisions

This plot demonstrates both daily and monthly differences in the distribution of pressure variable.

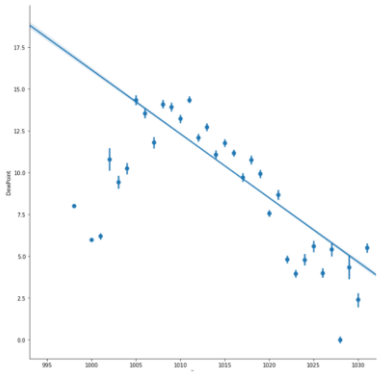


Figure 3.2.12.6 Linear model Plot of Dew Point and Pressure

Pressure and Dew Point have negative linear relationship.

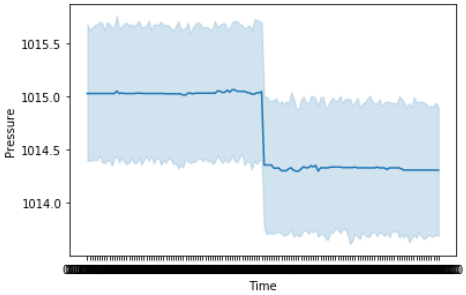
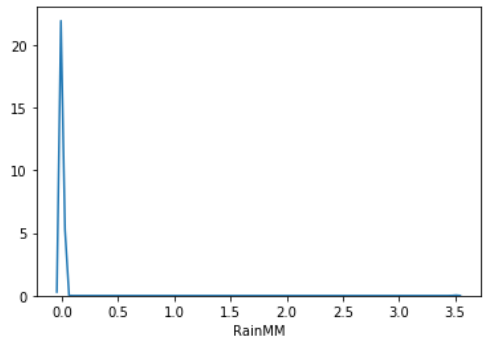


Figure 3.2.12.7 Time Series Plot of Pressure

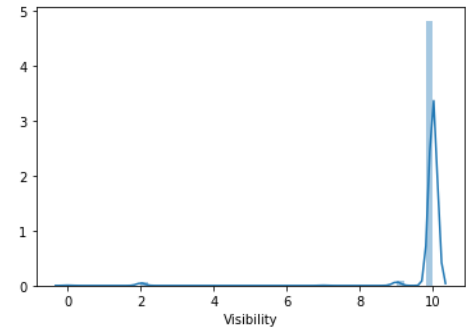
The plot shows 2 different seasons in first half of the year pressure is relatively high and after a drop it continues consistently until the end of the year.

3.2.13 RainMM and Visibility

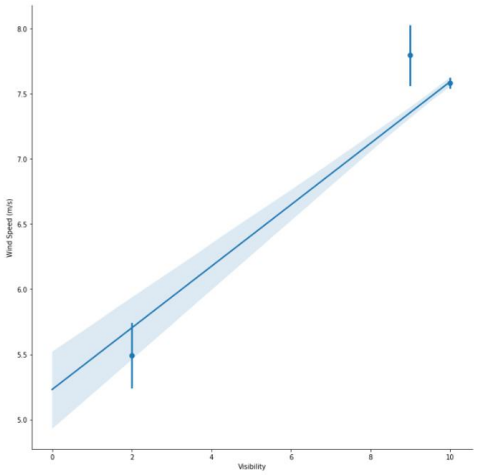
For RainMM and Visibility variables we could not obtaion meaningful distribution plots or histograms because these 2 variables stays same nearly on 90% of the time and because of that the distribution plots and histograms does not show meaningful results. For RainMM it stays as 0 nearly every time and because of that we could not create neither a histogram nor a meaningful linear model plot. We put here some examples to show that it is not meaningful to have plots for RainMM. The case is very similar in the visibility variable, it stays as 10 most of the time, we also put the plots for it to, to show that they do not make sense.



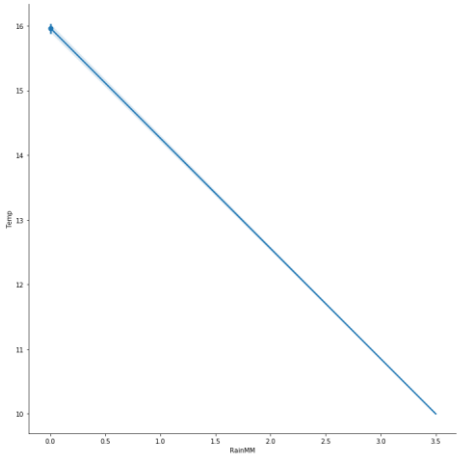
3.2.13.1 Distribution Plot of RainMM Attribute



3.2.13.2 Distribution Plot of Visibility Attribute



3.2.13.3 Linear Model Plot of Visibility and Wind Speed

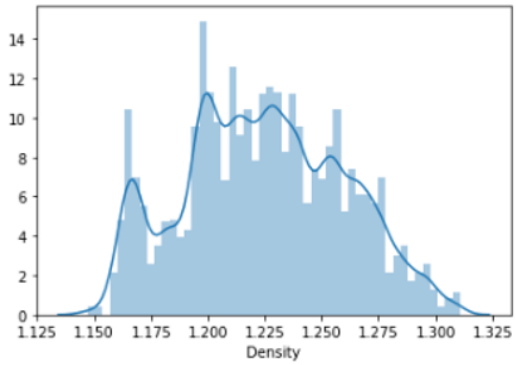


3.2.13.4 Linear Model Plot of RainMM and Temperature

As mentioned above plots for these 2 variable do not make sense.

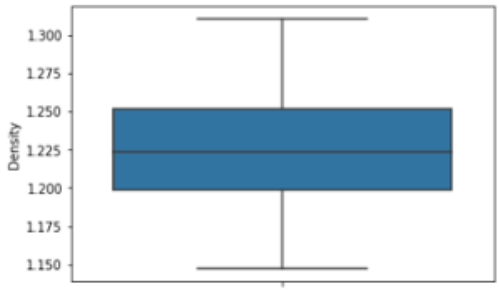
3.2.14 Density

Density a new variable but actually it is calculated with the help of temperature and pressure values. We added density as a new variable to our set because air density is a very meaningful term in theoretical calculations of electric production of wind turbines. We also observed this in our linear model plots between Theoretical Power Curve and density graphs. With these new variable we may use it directly to guess real production also instead of using temperature and pressure variables individually.



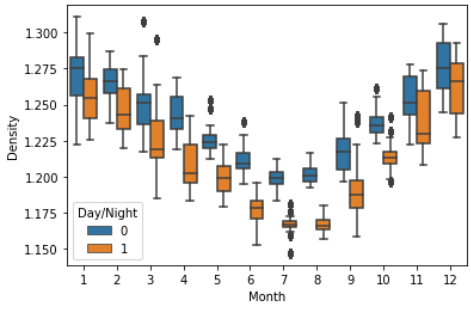
3.2.14.1 Histogram of Density Attribute

Our distribution seems a little similar to normal distribution, except 2 peaks at the left.



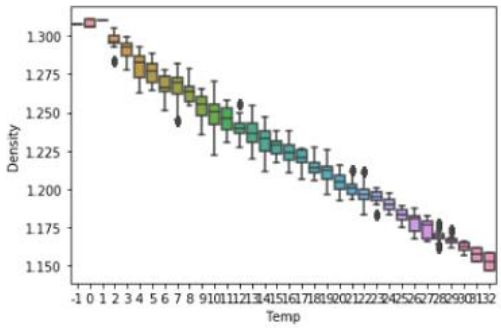
3.2.14.2 Boxplot of Density Attribute

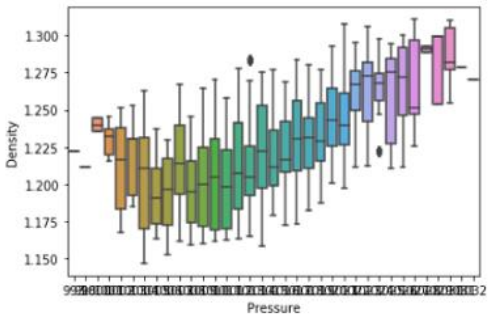
Boxplot shows us an equally distributed data, with no outliers.



3.2.14.3 Boxplot of Density Attribute with different subdivisions

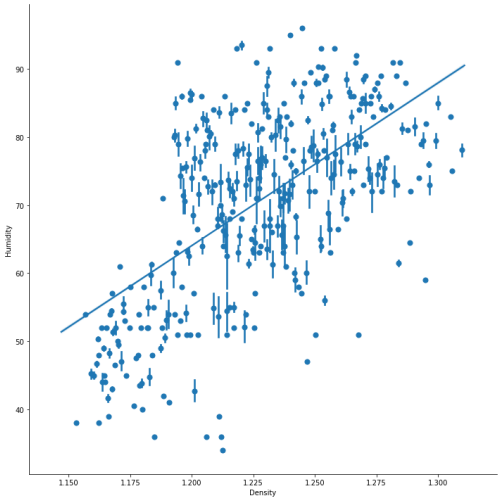
Since density is calculated with temperature and pressure, it is inherently have the temperature variable's nature. It differs in day-night and in monthly basis similar to the temperature. These dependency can be seen in the following 2 graphics.





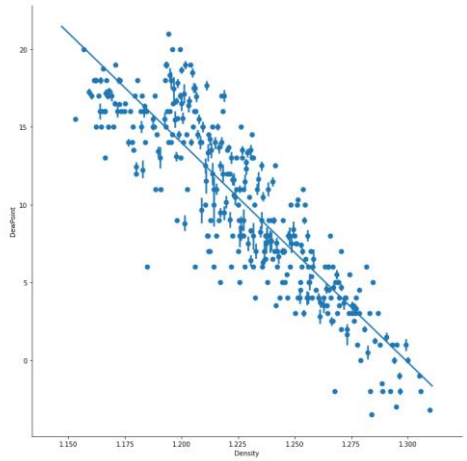
3.2.14.4 Boxplot of Density Attribute with respect to Temperature and Pressure

It can be seen that as temperature increases densities seem to decrease and their variance also decreases. Differently as pressure increase density and its variance increases in the boxplots.



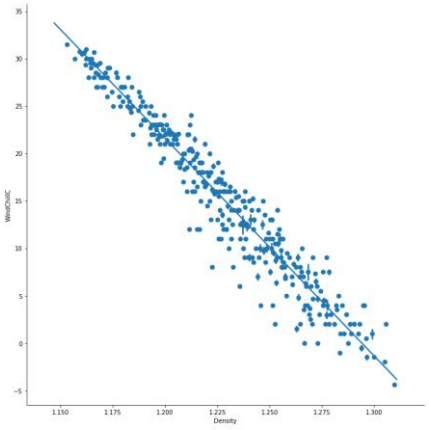
3.2.14.5 Linear model Plot of Density –Humidity Attributes

Density and Humidity have positive linear relationship.



3.2.14.6 Linear model Plot of Density –DewPoint Attributes

Density and Dewpoint have negative linear relationship.



3.2.14.7 Linear model Plot of Density –WindChill Attributes

Density andWindChill have negative linear relationship.

3.3 General Picture

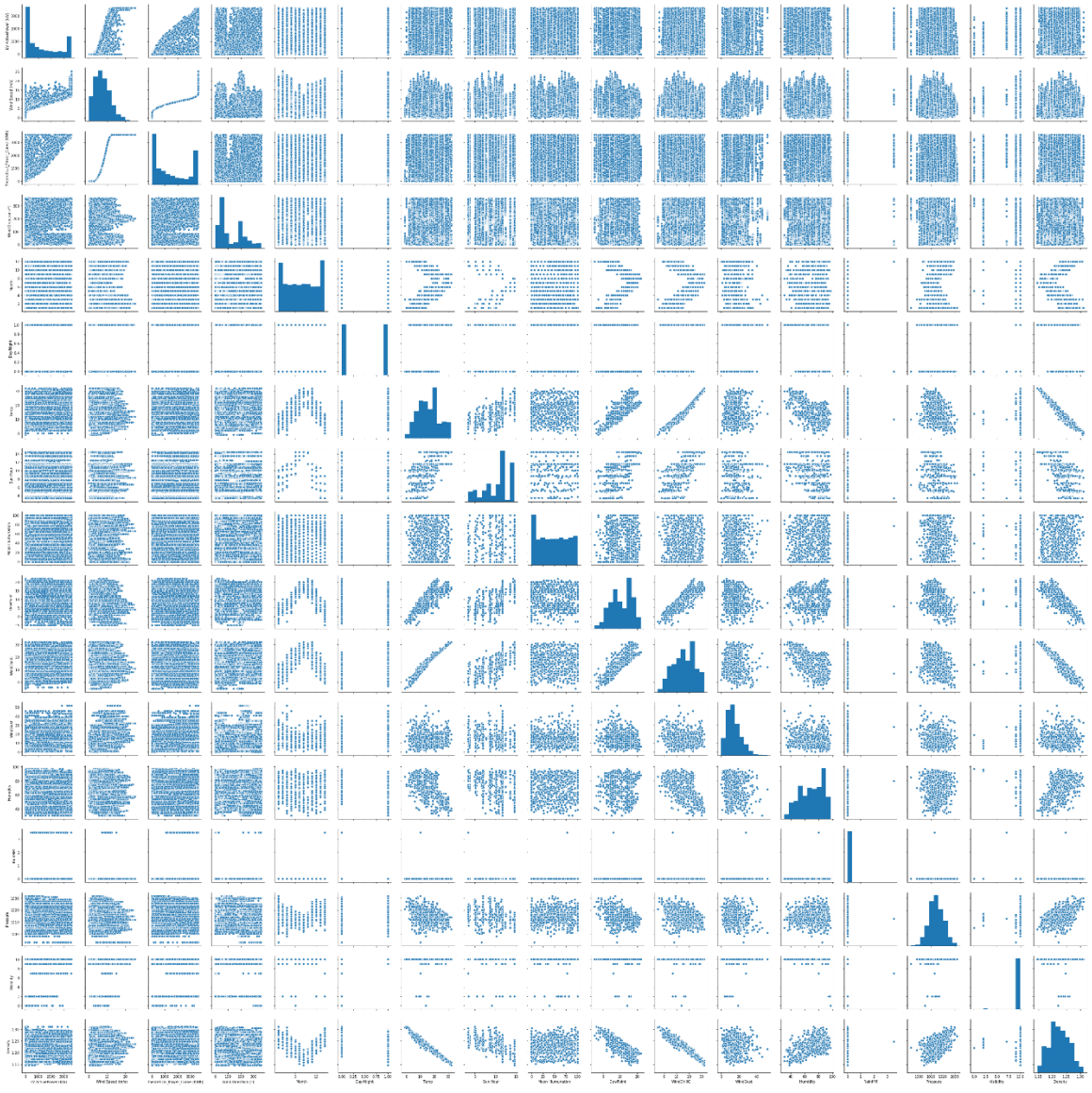


Figure 3.3.1 Scatterplot Matrix

Here in this scatter plot matrix we can see all possible combinations of relationships of variables, sadly because we have a dataset with nearly 50000 rows, the scatterplots are not clear. To overcome this difficulty we plotted our individual linear model plots with x_bins parameter. It puts the dataset in the specified bins to visualize graphic more clearly but this parameter does not work on the pairplot function we used to create this scatterplot matrix.

Still we could use this graphic to have a general understanding about distributions of our attributes.

3.4) Conclusion

In this part we examined our variables in a more detailed manner with the help of different plots. We started these analysis with aims of improving our understanding about our variables and discovering meaningful relationship between variables. After rigorous examinations, we can say that we achieved our goals.

The usage of histograms, distribution plots and boxplots helped us to understand the general behavior of our variables. In further analysis of our variables with dividing them into sub categories, we were able to see that, while in some variables there is no difference between day/night or month measurements for some variables nearly every month has its own unique distribution. These differences imply that we need to focus on those variables to gain further information about their situation.

The usage of linear model plots were crucial to discover linear relationship between variables. We used this plot and tried nearly every possible combination of two variables in our dataset. We discovered some expected and unexpected relationships between variables. For example, we discovered that Theoretical Active Power Generation variable behaves differently than the LV Active Power Generation variable. Theoretical Power Generation variable is a calculation to guess the real LV Active Power generation from the wind turbine. We expected these 2 variables to behave similarly but even though they behave similarly in general, Theoretical value has a stronger relationship with the temperature and density. We discovered this via the help of the linear model plots. These plots is directly related with our purpose since we want to develop regression models, it is very helpful to see the relationships between variables and types of these relationships for us to develop the right type of the regression models.

At the end, throughout exploratory data analysis phase, we tried to explore our dataset fully and after hard work, just like every explorer we got our reward. We developed a better and proper understanding of our dataset, we observed relationships between our variables, we examined distributions of our variables and drew meaningful conclusions. We will use this knowledge to build our predictive models to be better and more accurate.