**Descriptive Analysis**

Original time series data consist of hourly production values for dates from 2021-02-01 to two days before current date. Analyzing the data, we can see that there is an upper limit for the production values, of which production values cannot exceed. Construction of the production facility started in beginning of 2021. For this reason, capacity is small at the beginning of the data and increasing constantly until June 2021. Then, it is constant for some time around 40KWp. It decreases to 35KWp in the beginning of 2022 and stays on that level since. This maximum capacity change is a significant problem to deal with because linear regression works in an additive manner and it will find the optimal coefficients for an assumed maximum capacity, which will not work well with capacity change.

We are also given a weather data, which consists of previous weather values and future forecasts for some categories. These categories are Cloud Low Layer, Relative Humidity, Temperature and DSWRF, each are given their hourly values for 9 different locations. Values of a specific category for different locations are highly correlated because locations are close to each other. Therefore, using values for all locations is not necessary, and might lead to a possible overfitting.

Importance of values from each location might differ based on the hour of the day we are aiming to predict. For example, cloud low layer value at facility coordinate is important for hour 12. However, at hour 16, angle of the sunlight changes and sunlight reaching to the facility passes not from the clouds directly above the facility, but from the clouds around that area. Thus, cloud low layer value of another coordinate will be important for this hour.

Cloud low layer values and production values are negatively correlated. This is because if the sky is cloudy, sunbeams can’t pass through, therefore decreasing the production.

DSWRF values and production values are positively and significantly correlated. They will be useful to predict production values in our model.

Relative humidity values are negatively correlated with production. and temperature values are positively correlated with production.

Newly created smoothed production variable is highly correlated (=0.789) with the production value. This variable will be important for our prediction. Only problem is that this variable takes the mean of previous and future 21 days’ production values and fills the remaining values with the last day’s smoothed production value. If we want to predict values near a season change, this variable will create bias and result in an overprediction or underprediction.

There is no production at hours 0, 1, 2, 3, 4, 20, 21, 22 and 23. There is an abysmal production at hour 5 only for few days. For these reasons, and because all the hours have their unique patterns (hourly seasonality) hourly models are created for hours 6 to 19.

**Conclusion**

As a result of our work, we found that the most important variable for each model turned out to be the smoothed production. This is because recent production values are correlated to the tomorrow’s production values, and it is unlikely that tomorrow will have a drastically different values from near past. Other predictors were expected to be important, however they turned out to be not that important. This is probably because their effect is more complex and not easy to interpret.