IE360-Statistical Forecasting and Time Series Models
Group Project
Group-15

Salih Canbolat-2020402141
Aris Dikran Nişan-2020402075
Zeynep Betül Özarslandan-2020402048

1. **Introduction:**

In this project, it is expected to provide hourly solar power predictions of Edikli GES (Güneş Enerjisi Santrali) everyday for the next day. Edikli GES is located in Niğde at 38.29 North, 34.97 East (between 37.75-38.75° north latitude and 34.5 -35.5° east longitude.)

The assumption is that on day *d*, the predictions are needed for day *d+1* and the production values until the end of day *d-1* are known. There are 25 grid points nearby the power plant whose coordinates can be seen above Table.1.

	lat	lon
1	37.75	34.5
2	37.75	34.75
3	37.75	35
4	37.75	35.25
5	37.75	35.5
6	38	34.5
7	38	34.75
8	38	35
9	38	35.25
10	38	35.5
11	38.25	34.5
12	38.25	34.75
13	38.25	35
14	38.25	35.25
15	38.25	35.5
16	38.5	34.5
17	38.5	34.75
18	38.5	35
19	38.5	35.25
20	38.5	35.5
21	38.75	34.5
22	38.75	34.75
23	38.75	35
24	38.75	35.25
25	38.75	35.5

Table.1: The latitude and longitude pairs for weather variables

2. **Preparing the Data for Prediction:**

a. **Arranging the Data:** Two data sets are provided for being used in the prediction step, one of them is the daily production amount data and the other one is the regressors: the related latitude and longitude; the relative downward shortwave radiation flux (DSWRF); the relative total cloud cover data (in terms of percentage) for different type of clouds; the solar radiation related variables which can be related to the production level; the related categorical snow variable which takes 1 if it snows, 0 otherwise; and temperature data for twenty-five different locations.

Firstly, the data is turned into a wider format from the longer format based on the longitude and the latitude information, which makes the handling and analysis of the weather variables easier.

Additionally, the “production” and “whether info” data are joint regarding their date and hour.

b. Adding Dates and Simple Regressors:

The data is arranged based on date and hour, in addition; indices are added.
The maximum capacity in the past 24-hours starting from 48-hours before is needed and it is calculated by using the merged data.

c. Deleting Hours with No MaxCapacity Data:

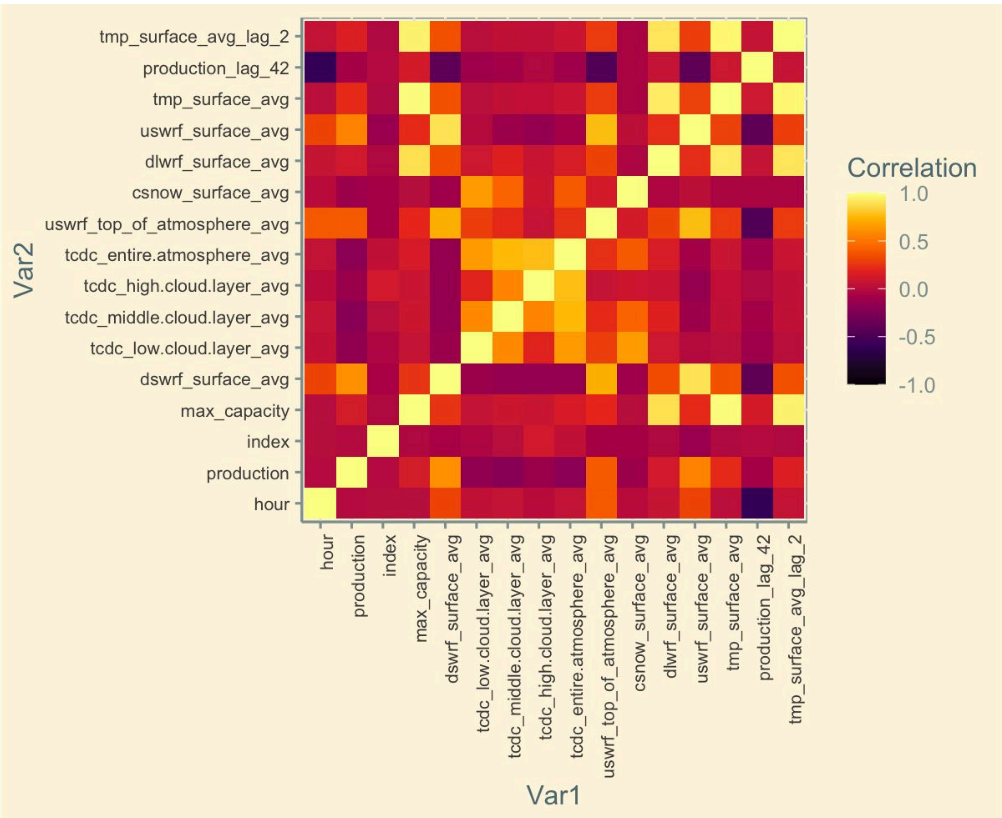
The hours which have no data as a result of the function calculate_max_capacity are removed.

d. Filling the Missing Data:

The production data from 24 hours before is added to every position where the data is missing.

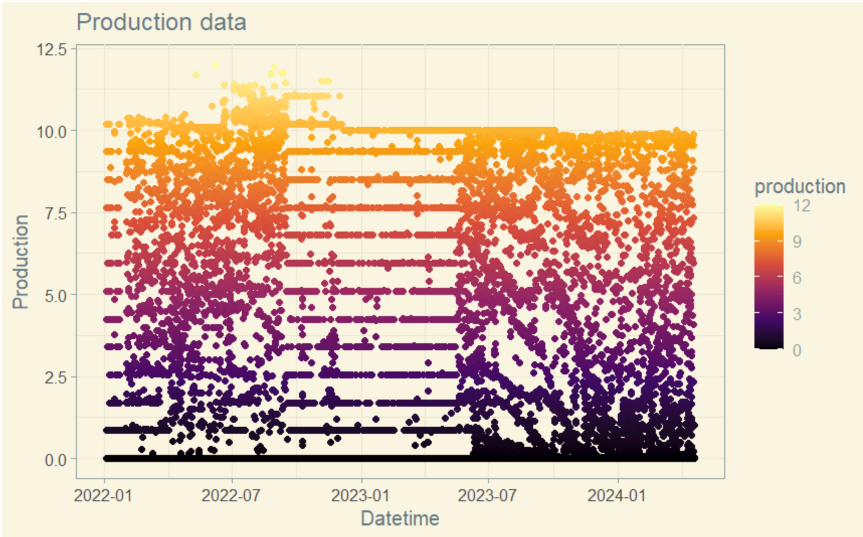
e. Adding Average of Regressors and Capacity Regressors:

The averages of each regressor for every 25 locations are calculated as new regressors. Directly adding only the average of the regressors could cause some logical problems since the capacity varies significantly from period-to-period and from hour-to-hour. Therefore, [Max Capacity*Average of Each Regressor] values are added as new regressors and the old ones are removed. The 42 hours lagged production data, 2 hours lagged temperature are also added as lagged regressors because the effect of temperature is not instant in real time, it takes some time for solar panels to get hot or cold.
With all these new regressors, indices and hourly production data, the correlation matrix is calculated and plotted as follows.



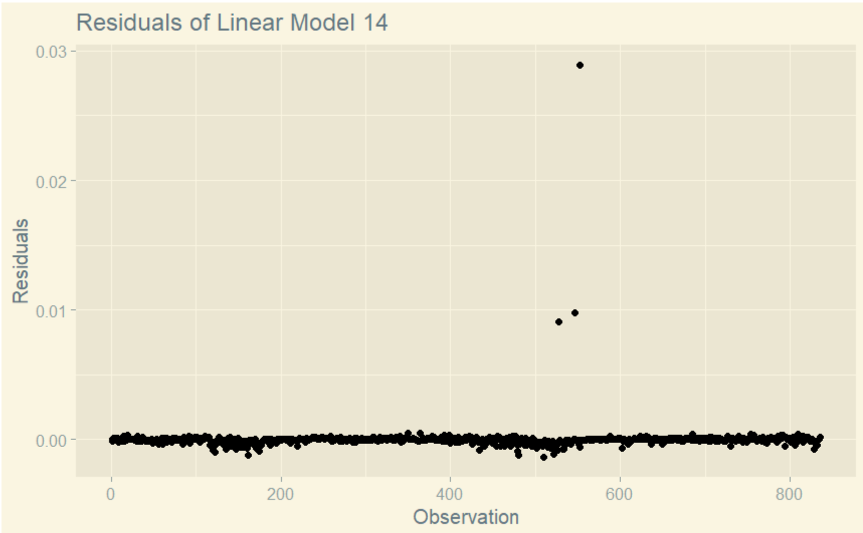
f. **Production Data Visualization:**

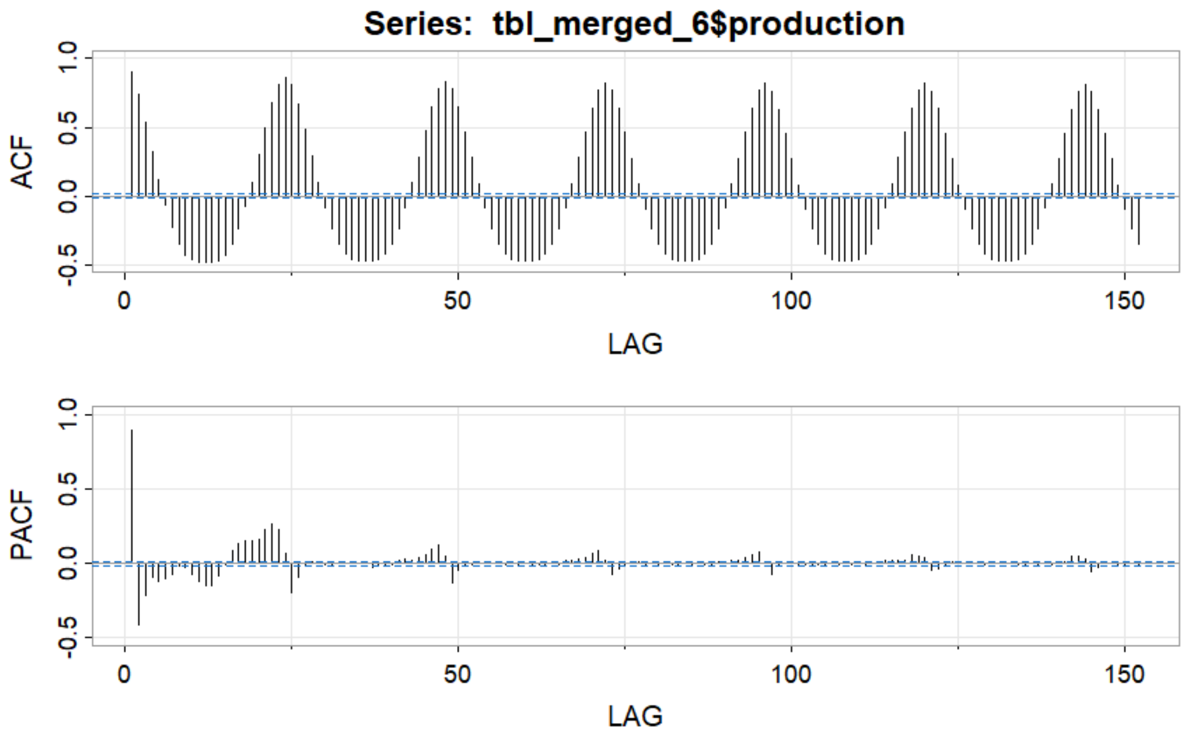
The production data is plotted to visualize the trend over time.



g. **Removing the hours with no production:**

Hours with no production are removed from the dataset, which are the hours between 20.00 and 05.00.

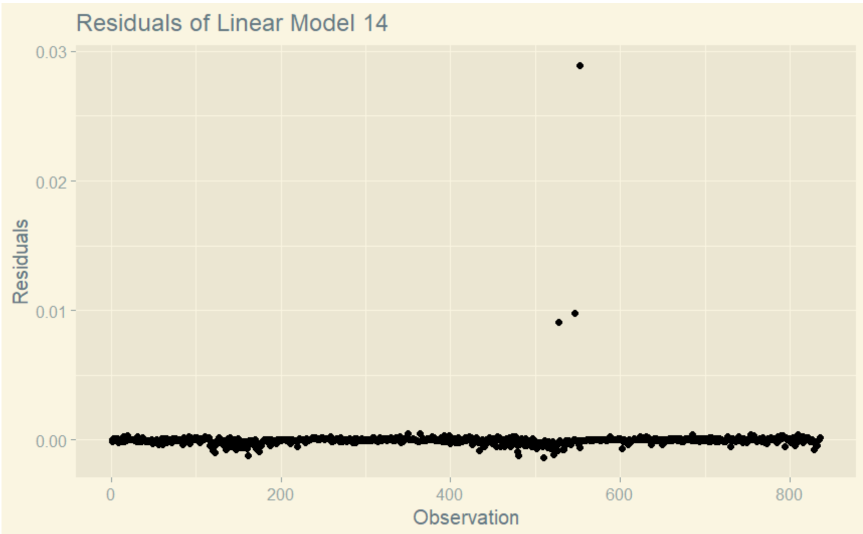




3. **Model Creation Phase:**

a. **Regression Model:**

The data is grouped by hour and the linear regression models are fitted for each one of them.
Residuals for each model are plotted to check whether the models are fitted significantly.

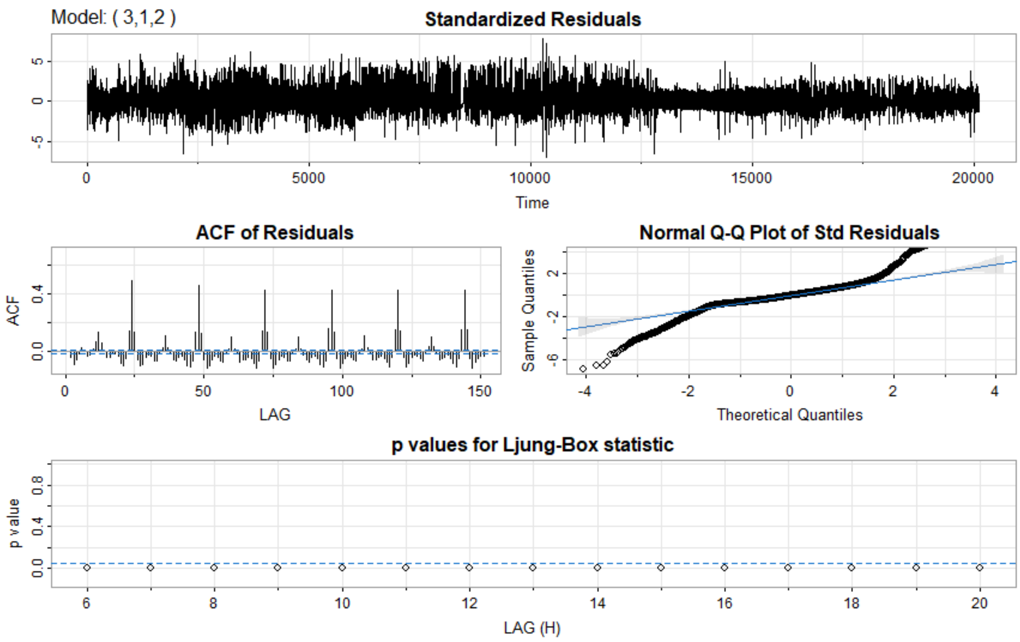


b. **ARIMA(3, 1, 2) Model:**

```r
# Auto ARIMA model
auto_arima_model <- forecast::auto.arima(tbl_merged_6$production)

# View the ARIMA model
auto_arima_model
```

In addition to the regression models, an ARIMA model is fitted for comparison by using the auto.arima() function. The ACF and PACF of it are checked and an auto-ARIMA model is created. Thanks to the 'astsa' package, a specific ARIMA model is fitted, which is ARIMA(3, 1, 2).

```
> auto_arima_model
Series: tbl_merged_6$production
ARIMA(3,1,2) with drift

Coefficients:
         ar1      ar2     ar3      ma1     ma2   drift
      1.9548  -1.1273  0.1081  -1.8732  0.8772  0e+00
s.e.  0.0085   0.0151  0.0077   0.0052  0.0054  6e-04

sigma^2 = 1.813:  log likelihood = -34511.86
AIC=69037.72   AICc=69037.73   BIC=69093.08
```



```
Coefficients:
         Estimate     SE   t.value p.value
ar1        1.9548 0.0085  230.5541  0.0000
ar2       -1.1273 0.0151  -74.5579  0.0000
ar3        0.1081 0.0077   13.9739  0.0000
ma1       -1.8732 0.0052 -362.8756  0.0000
ma2        0.8772 0.0054  161.8589  0.0000
constant   0.0000 0.0006    0.0191  0.9848

sigma^2 estimated as 1.812543 on 20101 degrees of freedom

AIC = 3.433517  AICc = 3.433517  BIC = 3.43627
```

c. **SARIMA Model:**

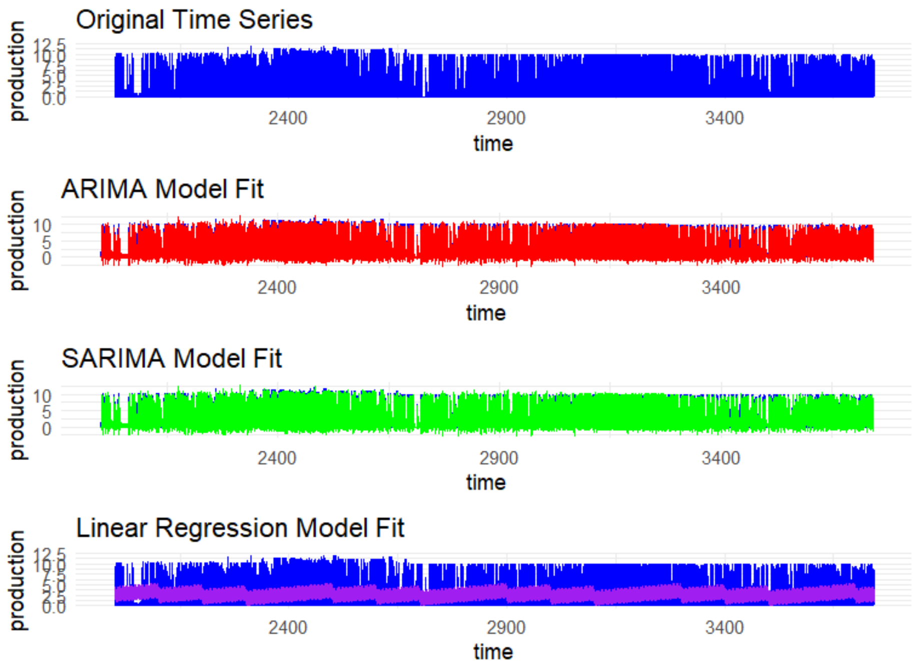
```r
# Replicate the model ARIMA(3,1,2) in astsa::sarima

# Check for missing values
sum(is.na(tbl_merged_6$production))

# Plot the production data to identify outliers
ggplot(tbl_merged_6, aes(x = index, y = production)) +
  geom_line() +
  labs(title = "Production Data", x = "Index", y = "Production")

# If necessary, apply a transformation
tbl_merged_6 <- tbl_merged_6 %>%
  mutate(log_production = log(production + 1))  # Add 1 to avoid log(0)

# Refit the SARIMA model with transformed data
sarima_model <- astsa::sarima(tbl_merged_6$log_production, p = 3, d = 1, q = 2)
```

Additionally, a SARIMA model is fitted in order to take into account any possible seasonal behavior of the production data. The fitted model is SARIMA(3, 1, 2)(0, 1, 1)[24], which integrates both the seasonal and the non-seasonal components of the data.

```
Coefficients:
         Estimate     SE   t.value p.value
ar1        1.9187 0.0079  244.0315  0.0000
ar2       -1.0490 0.0144  -72.9058  0.0000
ar3        0.0640 0.0075    8.5695  0.0000
ma1       -1.8886 0.0042 -446.5961  0.0000
ma2        0.8917 0.0044  201.8773  0.0000
constant   0.0000 0.0001    0.1483  0.8821

sigma^2 estimated as 0.123301 on 20101 degrees of freedom

AIC = 0.7456896  AICc = 0.7456898  BIC = 0.7484429
```

4. **Evaluation Phase:**

When the ARIMA, SARIMA and Linear Regression models are compared, it can be seen that the AIC, AICc and BIC values are distinctly decreased, which means that there is a significant change in the goodness of the model because smaller AIC, AICc, BIC values are demanded.



5. Conclusion:

The residuals are plotted for a single model. The steps outlined in this report represent a comprehensive approach to predicting solar power production using historical production data and various weather variables. The linear models by hour and ARIMA model provide different perspectives and methodologies for forecasting. The inclusion of a SARIMA model allows for capturing seasonal patterns in the data. Further refinement and validation of these models could enhance the accuracy of predictions.