**BOĞAZİÇİ UNIVERSITY**

**SPRING 2024**

**IE 360: STATISTICAL FORECASTING AND TIME SERIES**

**Doç Dr. Mustafa Gökçe BAYDOĞAN**

# Solar Power Production - Project Report

**Group 22**

**Erdem Efe Peynirci – 2020402168**

**Emine Uyar – 2020402087**

**Furkan Fazıl Çelik – 2022402282**

## TABLE OF CONTENTS

## INTRODUCTION

To create a sustainable world, it is vital to use sustainable energy sources such as solar energy. In order to use solar energy production capacity with maximum benefit and manage the solar power distribution, it is necessary to obtain a good predictionevaluate different alternatives . In this project, we focus on predicting hourly solar power production for the Edikli Güneş Enerjisi Santrali (GES) located in Niğde, Turkey (between 37.75-38.75° north latitude and 34.5 -35.5° east longitude). Accurate forecasting of solar power is crucial for efficient energy management and trading in energy markets. Our objective is to develop robust forecasting models using historical weather data to predict the solar power output for the next day.

The raw data were obtained from Edikli GES. This data includes weather measurements for 25 grid points near the power plant, covering various metrics such as downward shortwave radiation flux (DSWRF), cloud cover (TCDC at different layers), temperature (TMP), and categorical snow presence (CSNOW), collected on an hourly basis from January 1, 2022, to May 12, 2024. Additionally, we have the hourly energy production data for the same period. A preliminary analysis of this data reveals a clear daily seasonality pattern: energy production peaks between 10:00 am and 4:00 pm and drops to near-zero values between 8:00 pm and 4:00 am, reflecting the natural solar exposure cycle. Various weather metrics exhibit different influences on energy production; for instance, cloud cover tends to have a more significant negative impact on production compared to temperature.

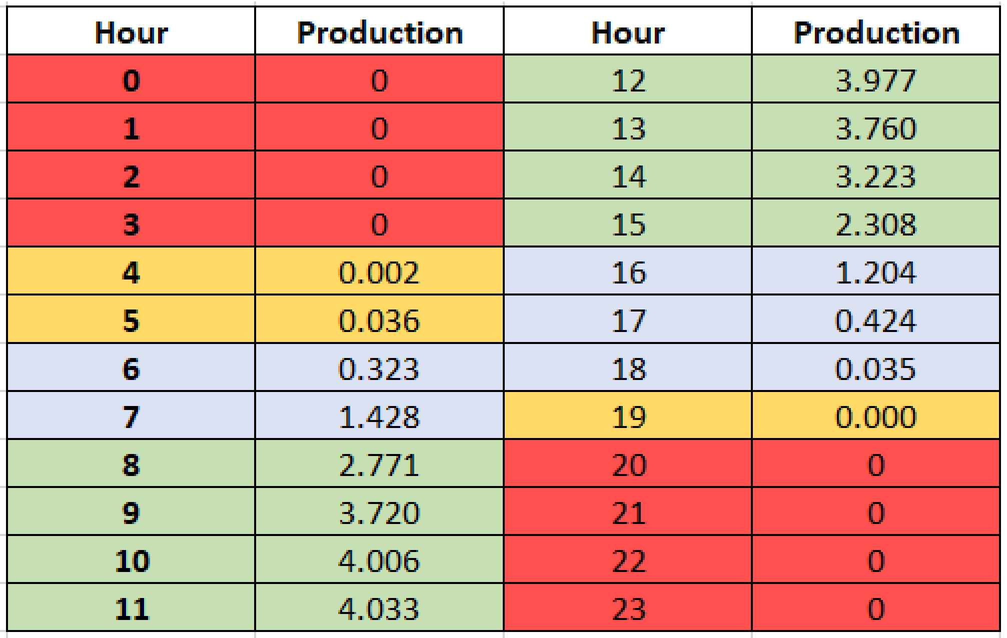
There are two datasets to be used in the project: one includes hourly production quantity data, and the other includes weather metrics (cloud cover, downward shortwave radiation, humidity, temperature, etc.) for different locations. We begin by reshaping the long-format weather data to a wide format. Then, we combine the reshaped weather dataset with the production dataset to form a comprehensive dataset for modeling.

After merging the datasets, we decided to build different approaches among the models discussed in the class and compare them based on evaluation metrics. To decide on which modeling methodologies we are going to use, we first trained a linear regression model and then we analyzed the ACF (Autocorrelation Function) and residual graphs to identify the issues in our current model. After that, we accordingly decided to apply ARIMA and Regression Tree methodologies and compare them to obtain the best possible solar power production predictions at the end of the process.

Before going into the details of our approach, it is crucial to emphasize that our observations on the dataset showed us the production amounts were similar within certain time intervals. To achieve more accurate results, we divided our dataset into different groups and tested all our models separately for each group. Also due to the fact that the maximum production amount allowed has changed to 10, we replaced any over this maximum limit to strictly 10.

## APPROACH

Before training and analyzing the models, the first step was to transform the hourly data of our variables, recorded at 25 different coordinate points, from long format to wide format. This transformation was essential to merge the production dataset with the relevant weather information dataset effectively. After merging the two datasets, we examined the average hourly production data, which below table demonstrates the results, to facilitate understanding and interpretation of the hourly production values.



| Hour | Production | Hour | Production |
|---|---|---|---|
| **0** | 0 | 12 | 3.977 |
| **1** | 0 | 13 | 3.760 |
| **2** | 0 | 14 | 3.223 |
| **3** | 0 | 15 | 2.308 |
| **4** | 0.002 | 16 | 1.204 |
| **5** | 0.036 | 17 | 0.424 |
| **6** | 0.323 | 18 | 0.035 |
| **7** | 1.428 | 19 | 0.000 |
| **8** | 2.771 | 20 | 0 |
| **9** | 3.720 | 21 | 0 |
| **10** | 4.006 | 22 | 0 |
| **11** | 4.033 | 23 | 0 |

Our analysis revealed that the average hourly production could be categorized into distinct groups throughout the day. The average production data clearly indicated four different time period groups:

**Group 1:** Hours when the sun is most effective, causing production values to reach their maximum limits (indicated in green on the table). This set encompasses the hours from 8 AM to 3 PM.

**Group 2**: Hours when the sun is either rising or setting, resulting in very low production values. (indicated in yellow on the table). This period occurs from 4 AM to 5 AM and at 7 PM. However, using a heuristic approach, 7 PM is excluded from this list because the prediction results for this hour have consistently been 0. Also, in the given data, the production values were strictly 0 before a specific date. So we excluded this period from the data and used this narrowed information for predictions.

**Group 3:** Hours when the sun is not yet at its peak, but considerable production is observed. (indicated in blue on the table). This set includes the hours from 6 AM to 7 AM and from 4 PM to 6 PM.

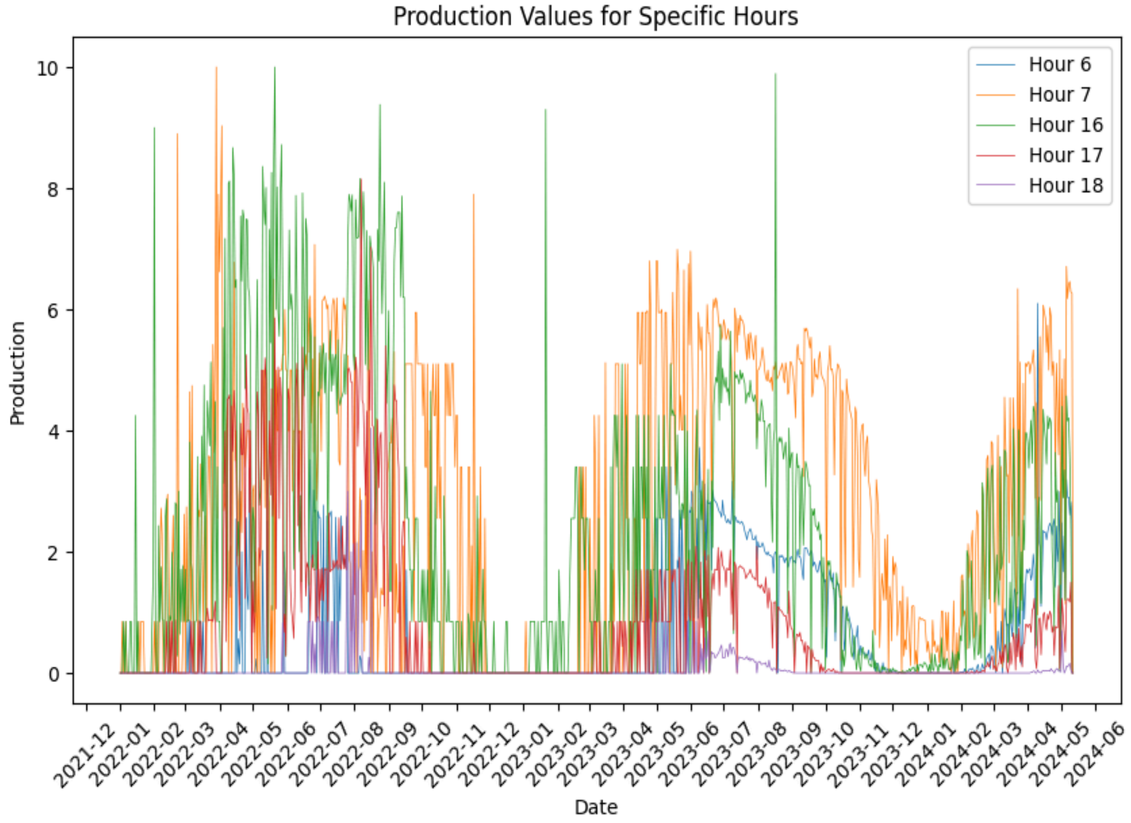
**Group 4**: Hours when the sun is absent and production is consistently zero (indicated in red on the table). It is obviously unnecessary to create a prediction model for this period, and it is fair to assign a value of 0 for all the predictions.

These observations helped us identify and understand the underlying patterns in solar power production, which are crucial for building accurate predictive models. By categorizing the data into these groups, we were able to tailor our models to capture the nuances of each time period effectively.

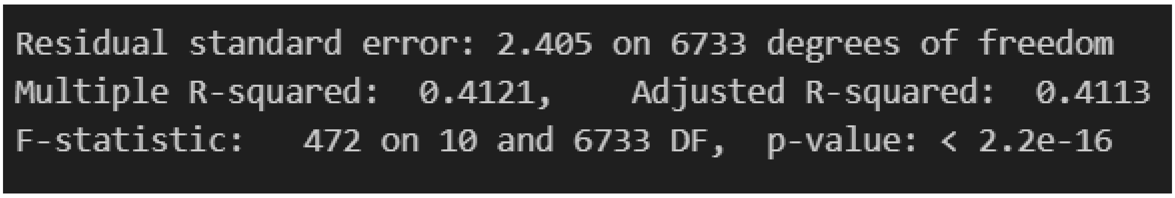
Moreover, the actual data given to us includes some values more than the current maximum production restriction of 10. So in the models described below, all values exceeding this limit was replaced with 10 to avoid obtaining erroneous predictions.

## MODEL 1: LINEAR REGRESSION

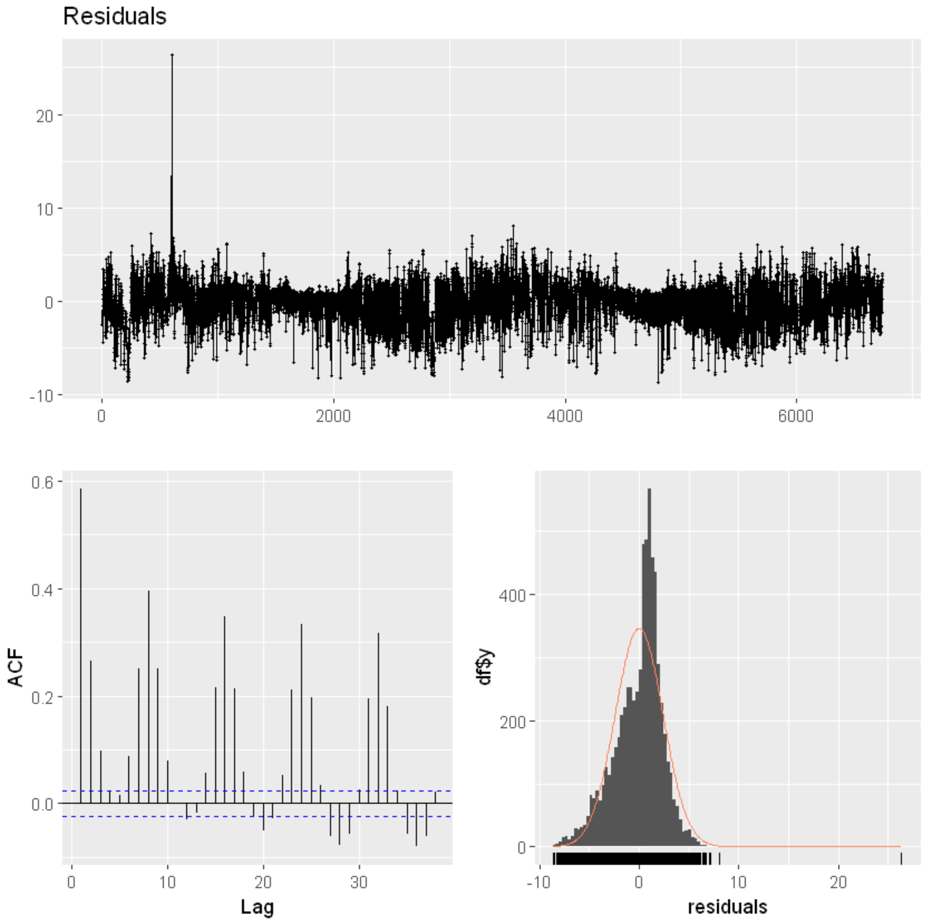
As the first model approach, a standard linear regression method is chosen. First, to observe the behavior of the data, we visualized the production amounts for some selected hours.



We constructed the model using all the independent variables as they were provided. Then, based on the outputs obtained, we conducted a residual analysis and observed the results below. Since all the time groups are modeled equivalently, only the results for Group 1 (8, 9, 10, 11, 12, 13, 14, 15) is given and the same procedure will be used for Model 1.

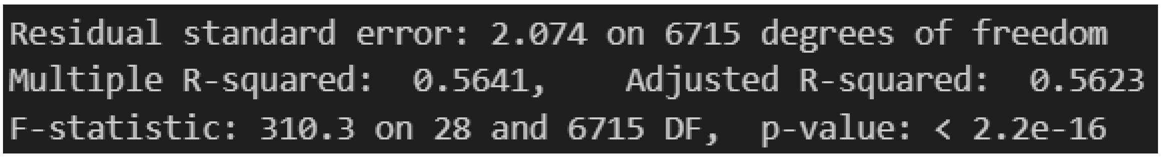


```
Residual standard error: 2.405 on 6733 degrees of freedom
Multiple R-squared:  0.4121,    Adjusted R-squared:  0.4113
F-statistic:   472 on 10 and 6733 DF,  p-value: < 2.2e-16
```

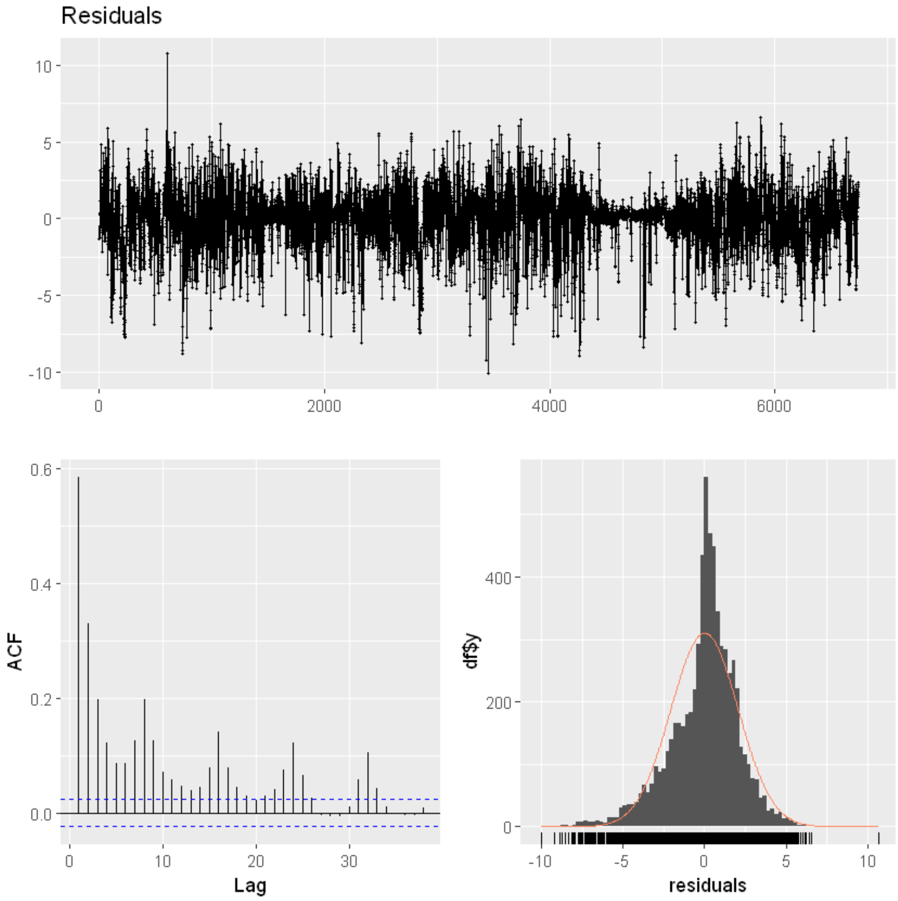


### Adding Dummy Variables

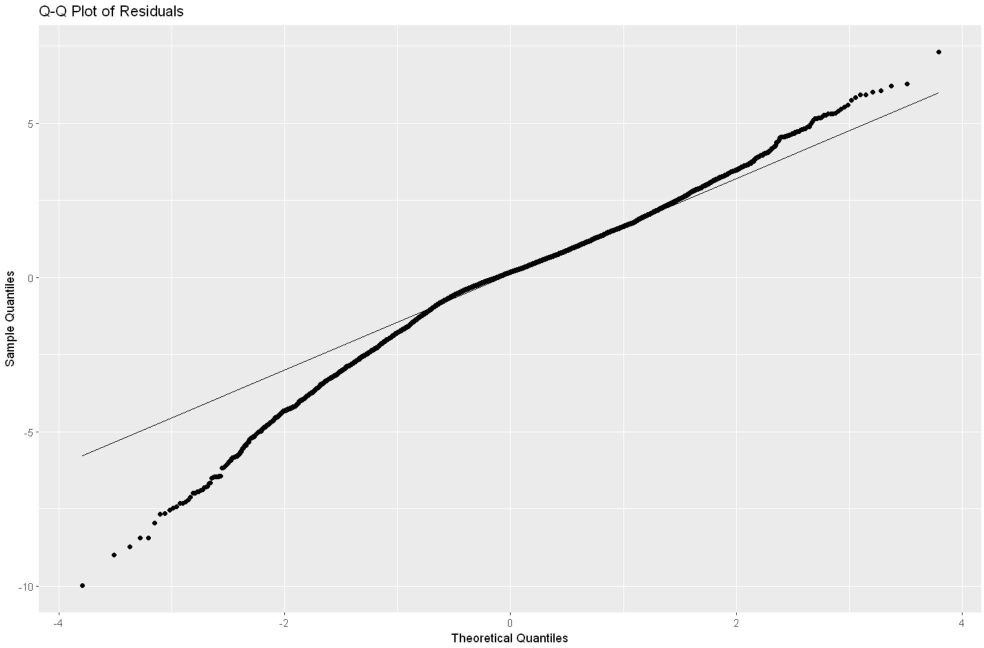
It is clear that our initial linear regression model contains a significant effect of seasonality. To integrate this effect into our model, we performed data manipulation by adding monthly and hourly dummy variables. The performance metrics, residual analysis and QQ plots of our enhanced model for each hour group are given below:



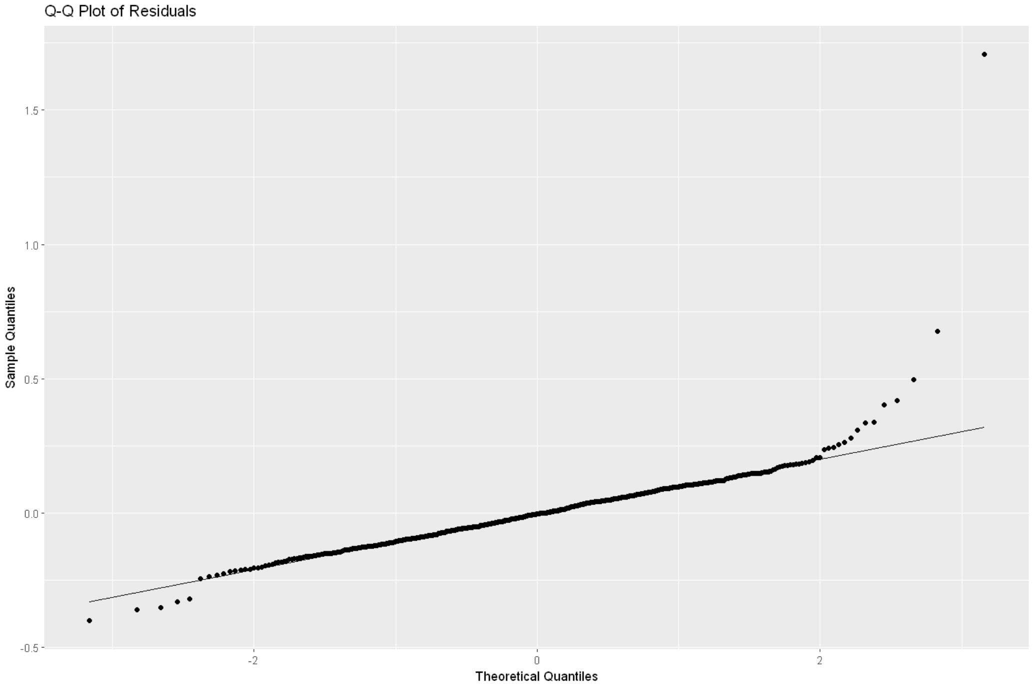
```
Residual standard error: 2.074 on 6715 degrees of freedom
Multiple R-squared:  0.5641,	Adjusted R-squared:  0.5623
F-statistic: 310.3 on 28 and 6715 DF,  p-value: < 2.2e-16
```



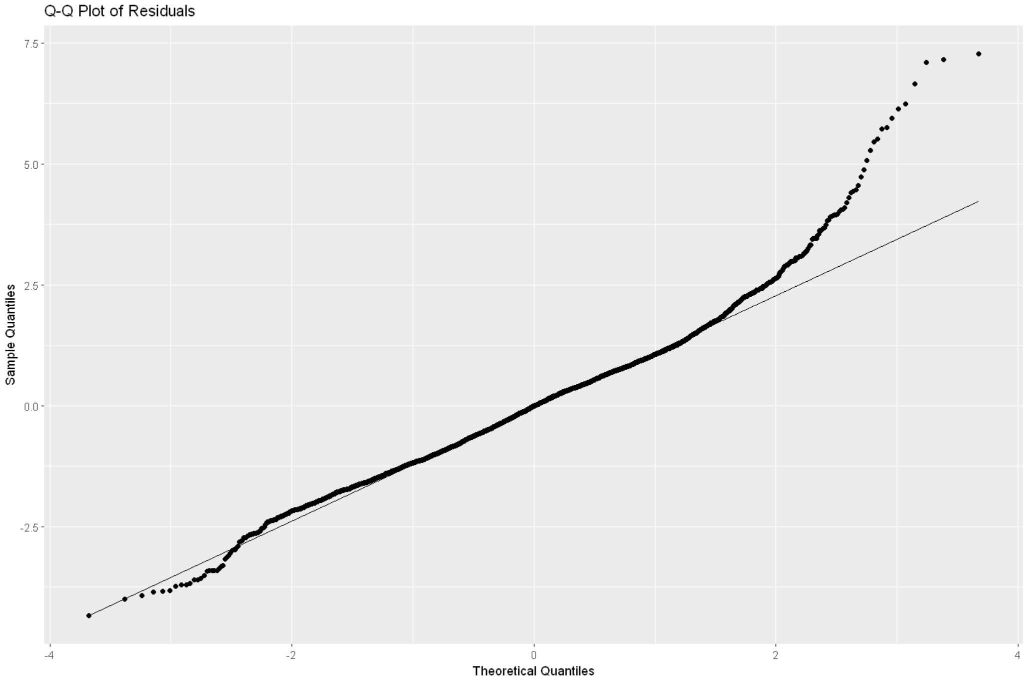
## Q-Q Plot of Group 1 (Hours 8, 9, 10, 11, 12, 13, 14, 15)



## Q-Q Plot of Group 2 (Hours 4, 5)



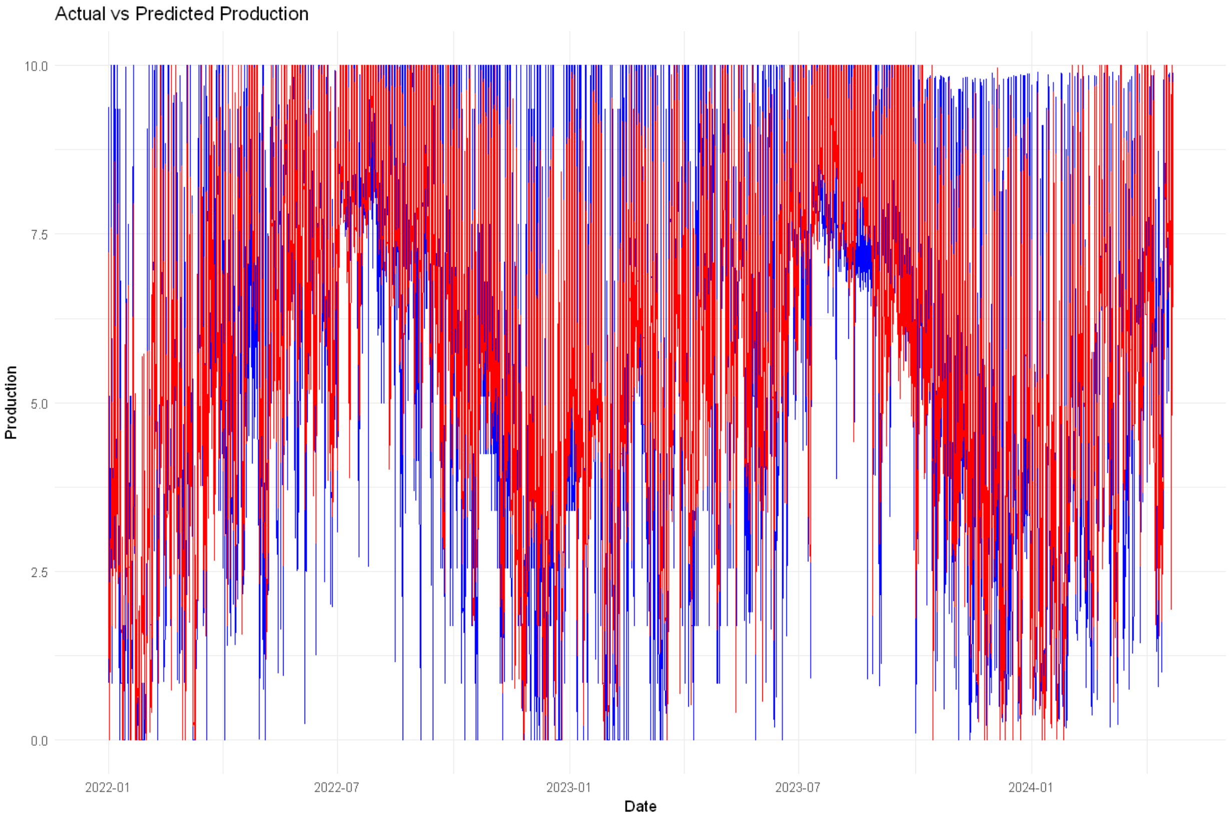
## Q-Q Plot of Group 3 (Hours 6, 7, 16, 17, 18)



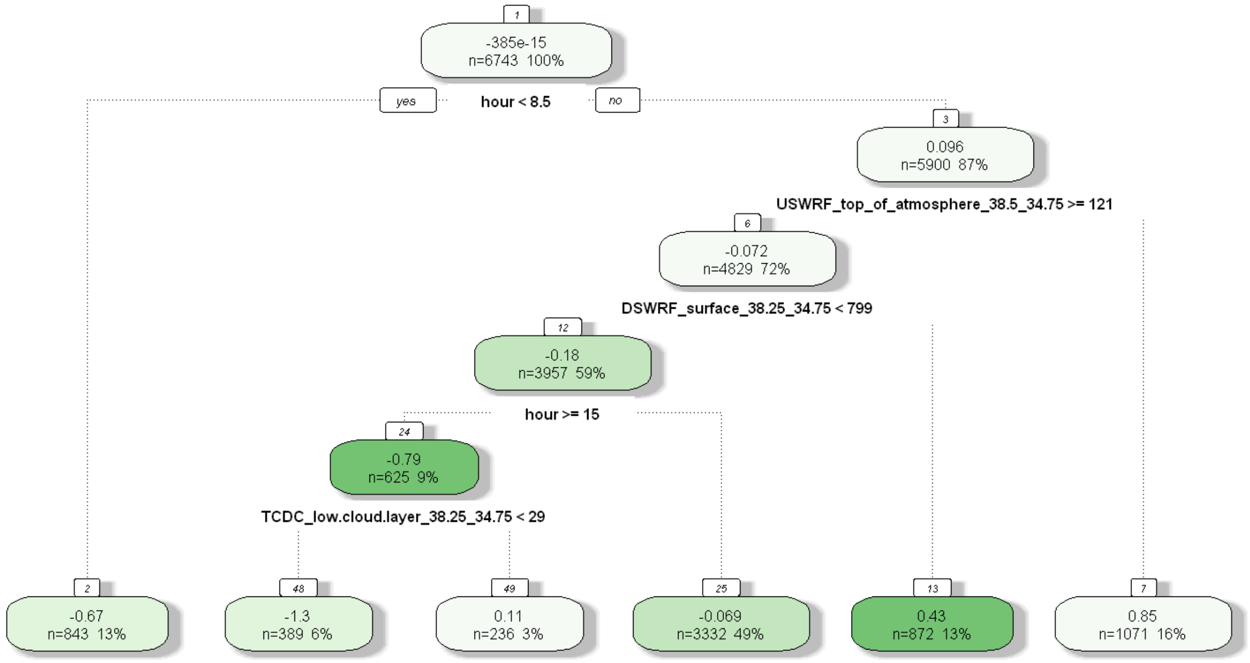
In the final LR (Linear Regression) model, after the manipulations, although there was an improvement compared to the initial version, we observed that the seasonality effect still remained significantly high and the performance measures were not at the sufficient levels. Therefore, new models that could provide more accurate results were needed.

# MODEL 2: REGRESSION TREE

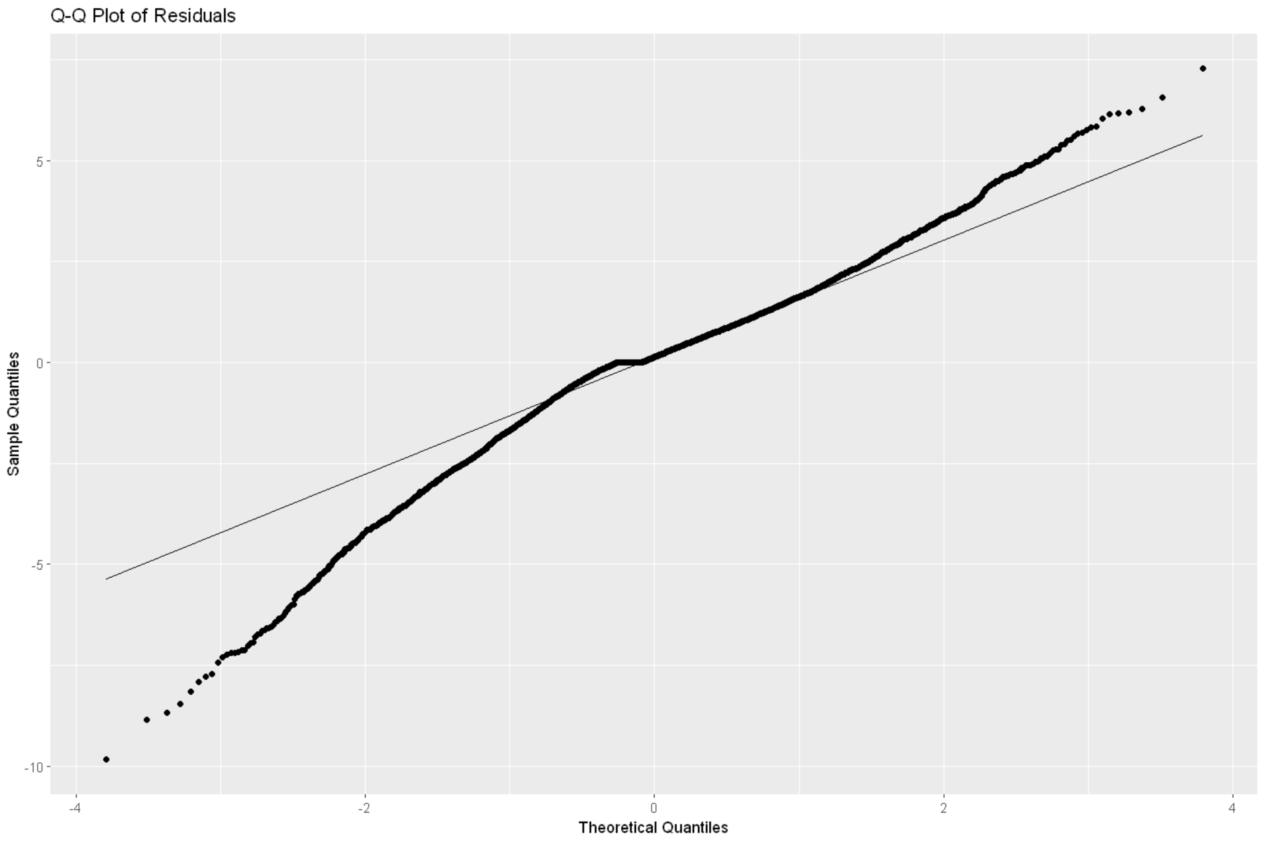
After evaluating the results of Model 1, we observed that the errors in our predictions were higher than the desired, resulting in relatively weak performance based on the evaluation measures. To be able to reduce the variance of the data by dividing it into nodes according to anova for improved results, we enhanced the performance of the model by integrating a Regression Tree model into the residuals obtained after the predictions of the linear regression model. This model narrows down the number of independent variable features by analyzing their significance on the results. Therefore the model complexity is decreased as well as its tendency towards overfitting. The visualization of predictions against actual value for group 1, obtained trees and QQ plots for each group are given below:



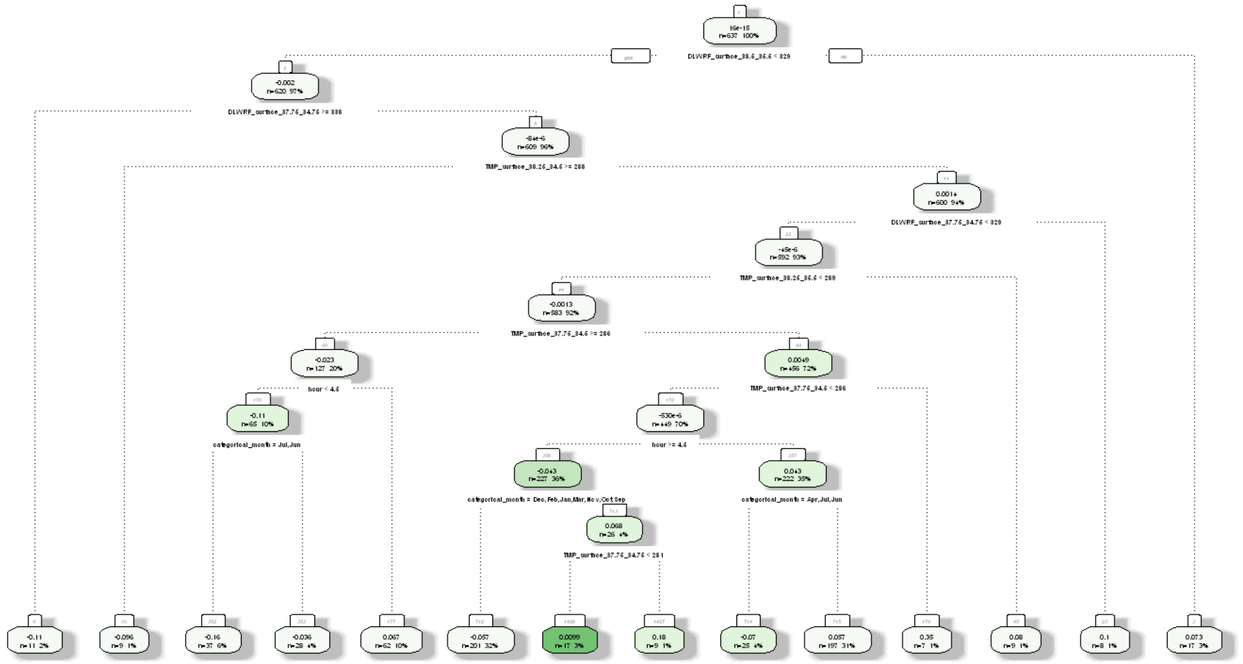
## Group 1 (Hours 8, 9, 10, 11, 12, 13, 14, 15)



Rattle 2024-Jun-05 17:30:07 Emine

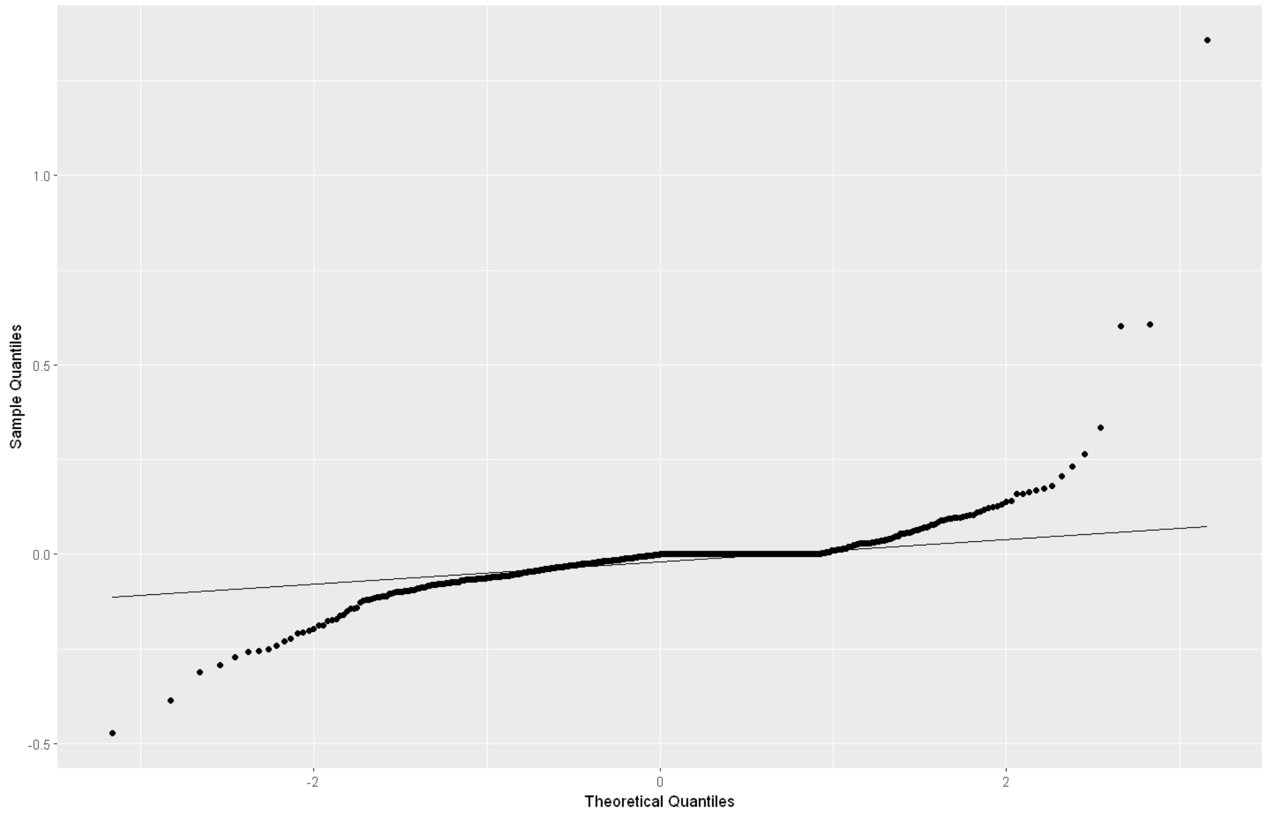


## Group 2 (Hours 4, 5)

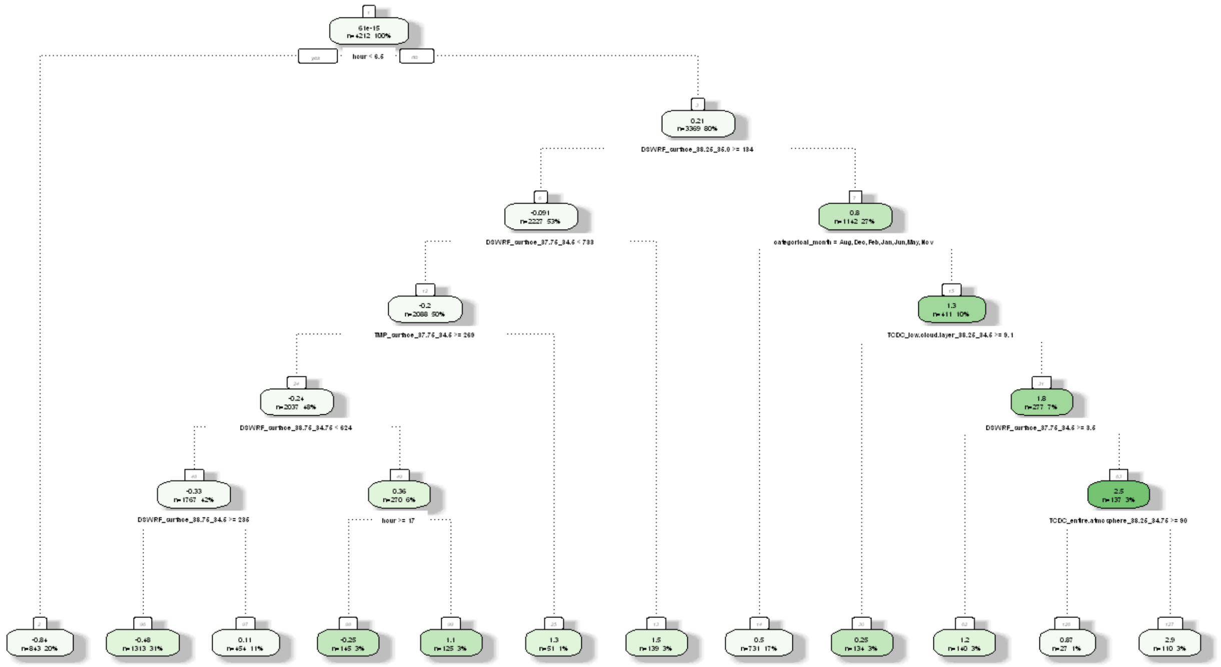


Rattle 2024-Jun-05 17:30:09 Emine

Q-Q Plot of Residuals

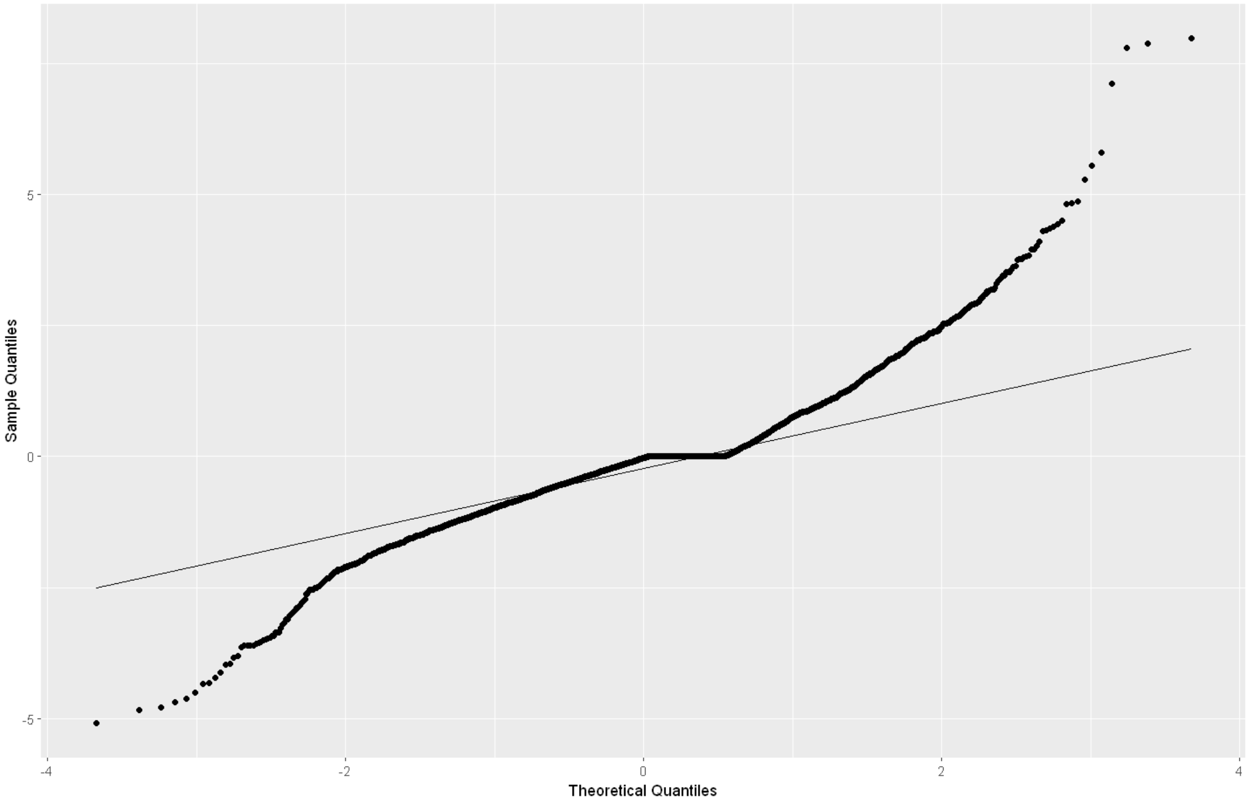


## Group 3 (Hours 6, 7, 16, 17, 18)



Rattle 2024-Jun-05 17:30:12 Emine

Q-Q Plot of Residuals



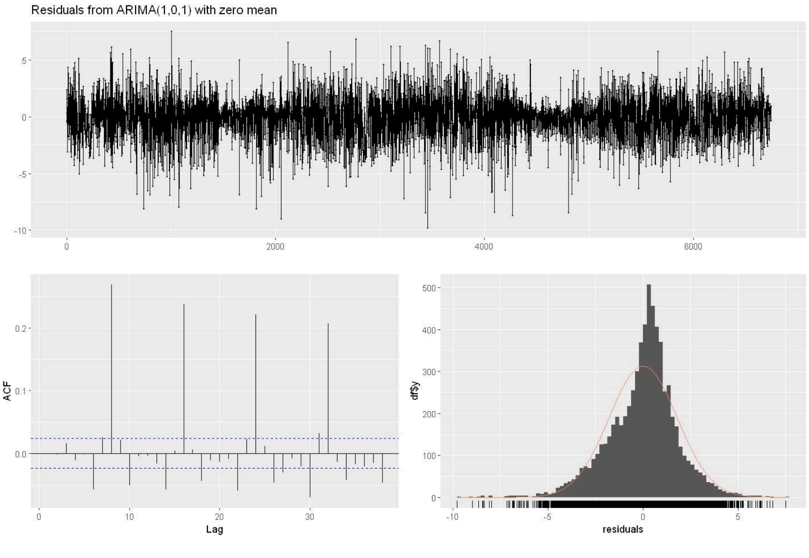
## MODEL 3: ARIMA APPROACHES

When evaluating the results of the first model, as seen in the graphs mentioned previously, although the mean of the residuals is close to 0, there was a large fluctuation range. Also, the residuals were still far away from being a white noise series. It means that there is still some information left that can be used in the model. Consequently, the ARIMA method was applied to the residuals in the second model to address this issue.

First, we used R's auto.arima function to obtain an initial model. Then, by examining the ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) graphs, we manually determined the order and parameters to create new ARIMA models. After comparing these two models, we selected the ARIMA model to be used for forecasting.

### Model 3.1. ARIMA with auto.arima()

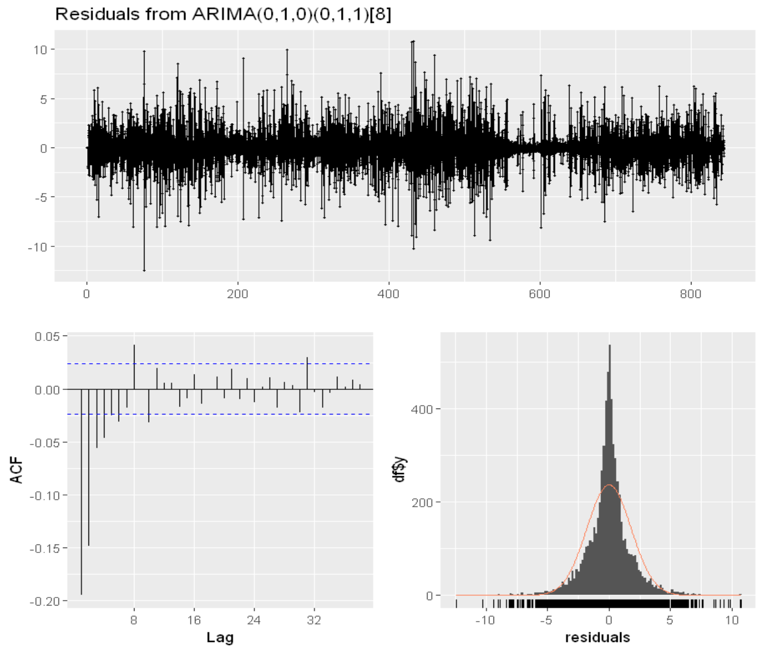
According to the auto.arima() function, the best parameter selection was (p,d,q) = (1,0,1). An arima model which uses these parameters created and it yielded the residual analysis below:

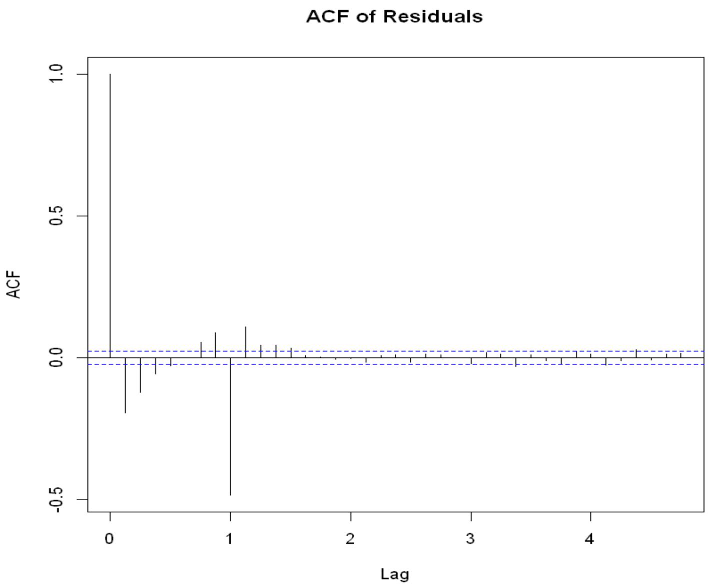


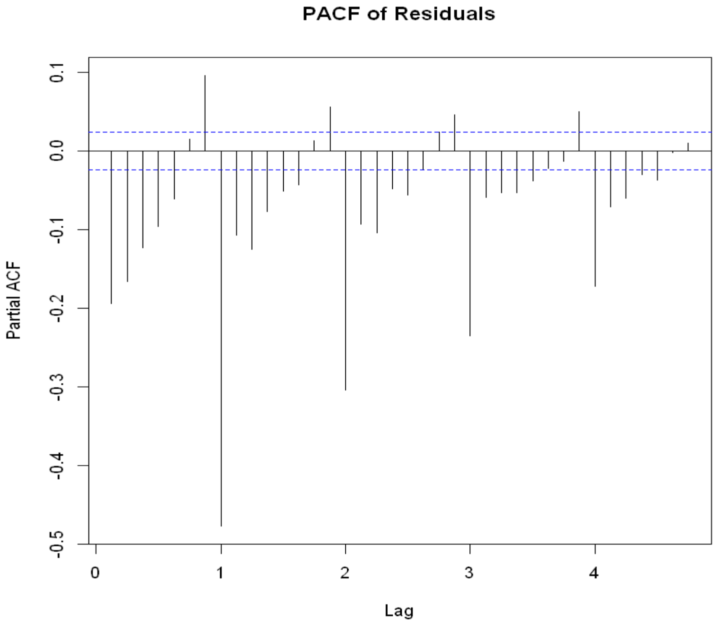
### Model 3.2. SARIMA (0,1,0),(0,1,1)[8]

We have used our theoretical backgrounds to find parameters for the ARIMA model instead of using the auto.arima function. A first order differencing and 8 seasonal differencing is applied to our data to obtain stationarity.

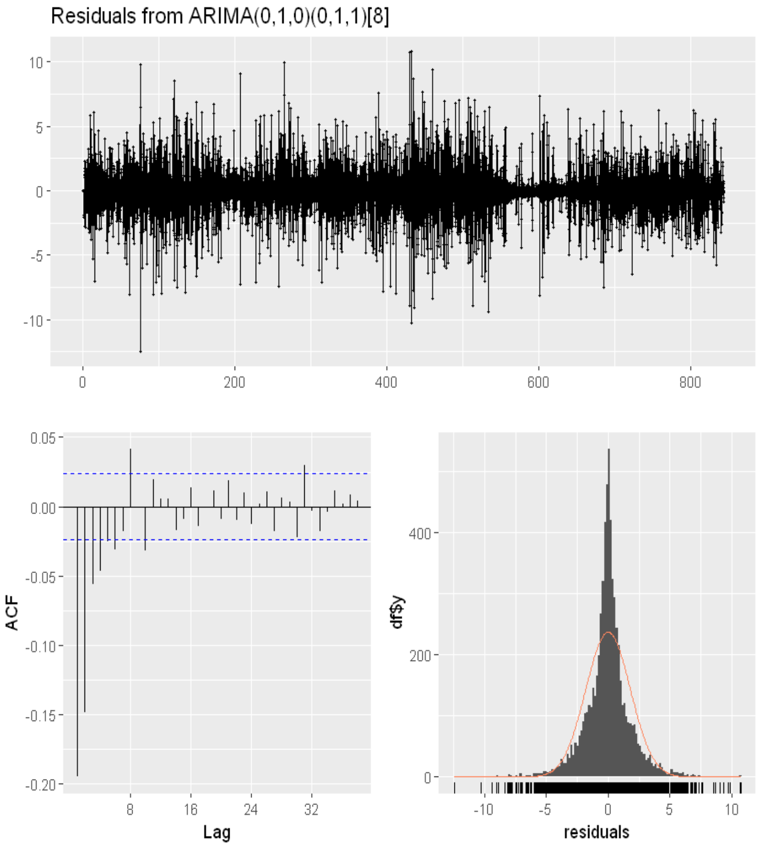
After that, the autocorrelation and partial autocorrelation plots have been taken into consideration to make the decision of the parameters. The resiudal analysis, autocorrelation and partial autocorrelation functions are as follows:

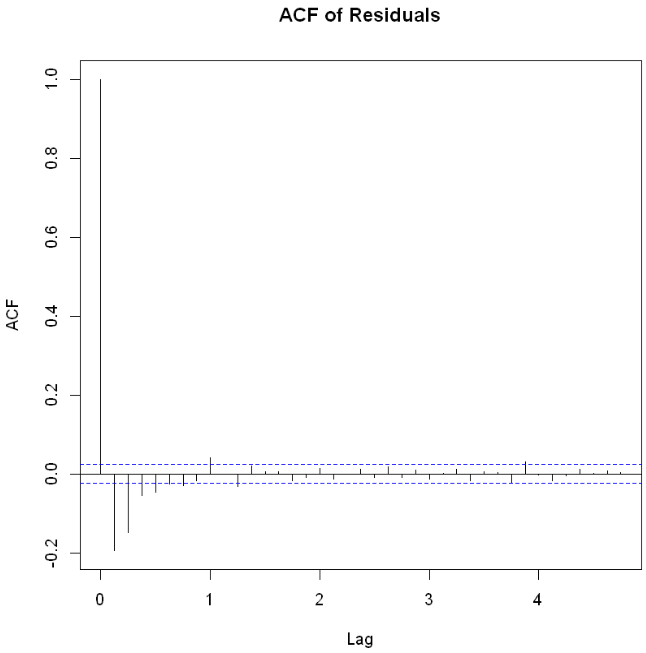


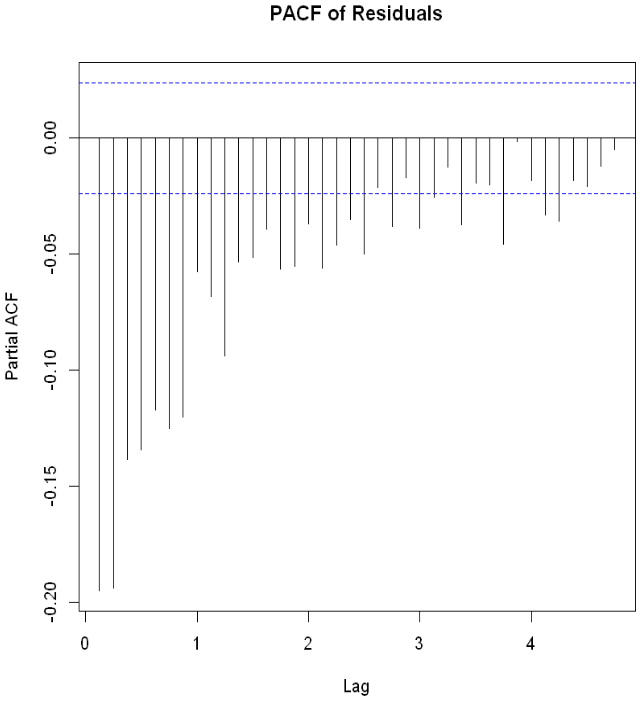




When the autocorrelation and partial autocorrelation plots are investigated, it can be seen that the autocorrelation function spikes in lag s (one season) and there is a spike in each seasonal period (s, 2s, 3s) on the partial autocorrelation function. Therefore, an SMA(1) model was used. Residual analysis and ACF/PACF plots are given below;

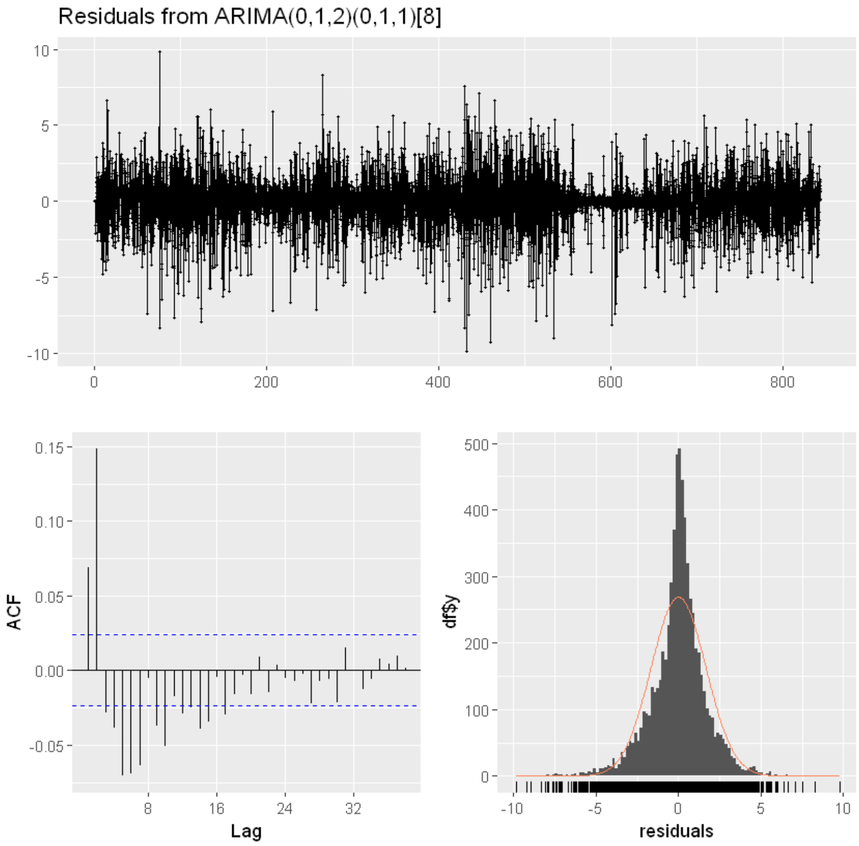


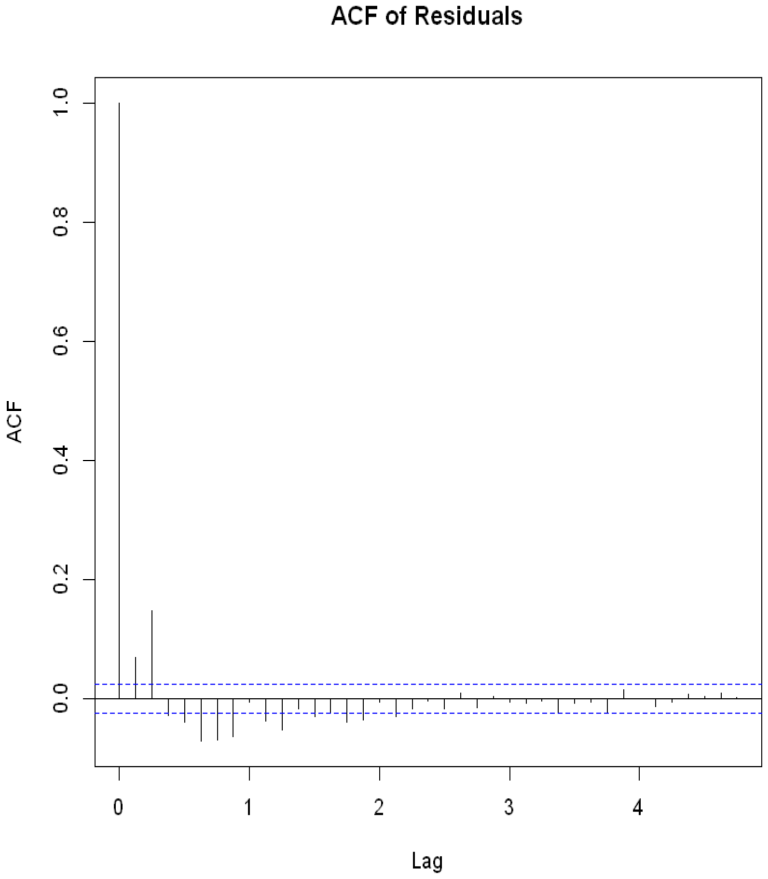


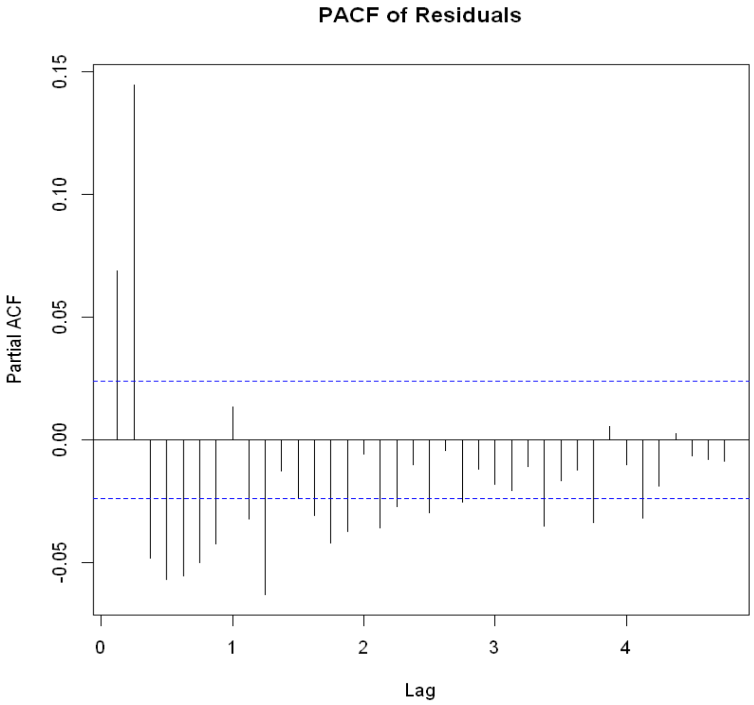


## Model 3.3. SARIMA (0,1,2)(0,1,1)[8]

Upon examining the ACF/PACF graphs of Model 3.2, an exponential decay in the PACF graph and a spike at lag 2 in the ACF graph were observed, leading to the application of a new model by using MA(2) approach. The residual analysis and ACF/PACF plots are given below;

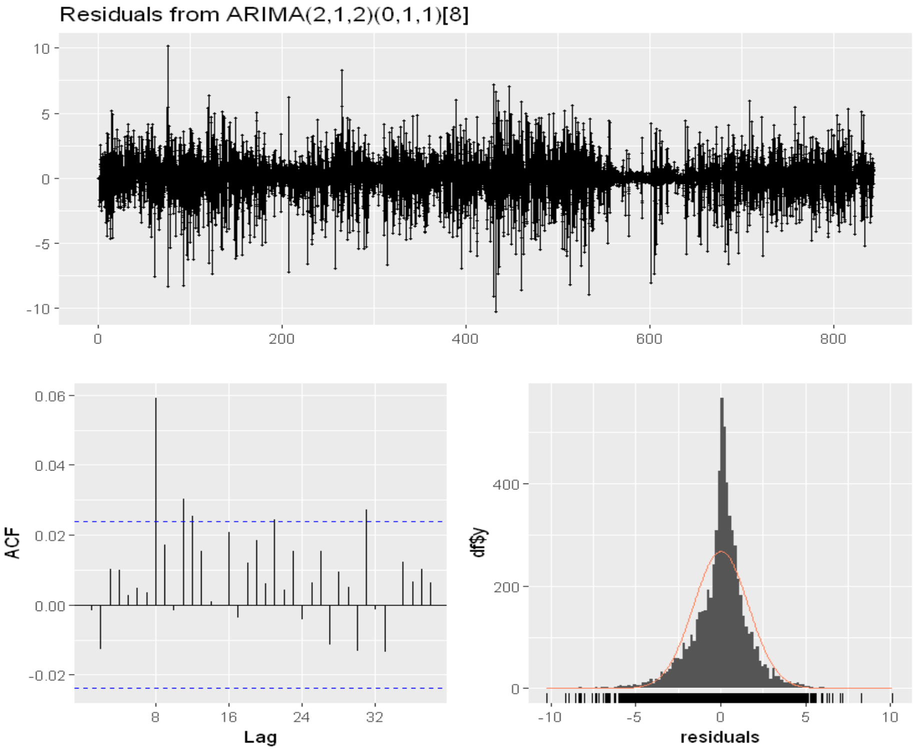

Residuals from ARIMA(0,1,2)(0,1,1)[8]
10
5
0
-5
-10
0
200
400
600
800
0.15
0.10
0.05
0.00
-0.05
ACF
8
16
24
32
Lag
500
400
300
200
100
0
df$y
-10
-5
0
5
10
residuals

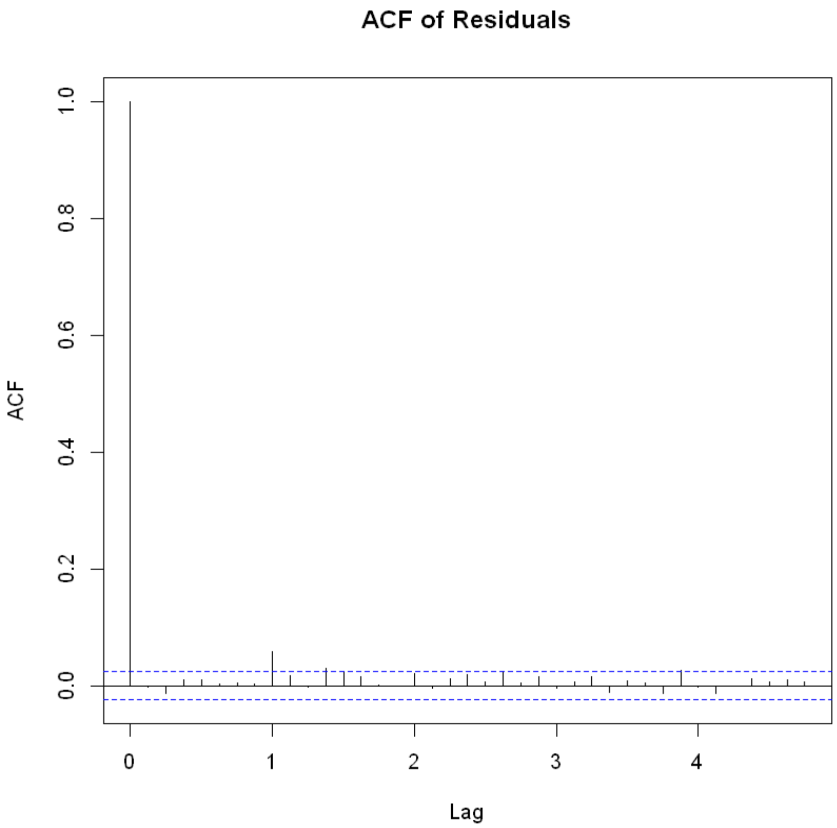

ACF of Residuals
1.0
0.8
0.6
0.4
0.2
0.0
ACF
0
1
2
3
4
Lag

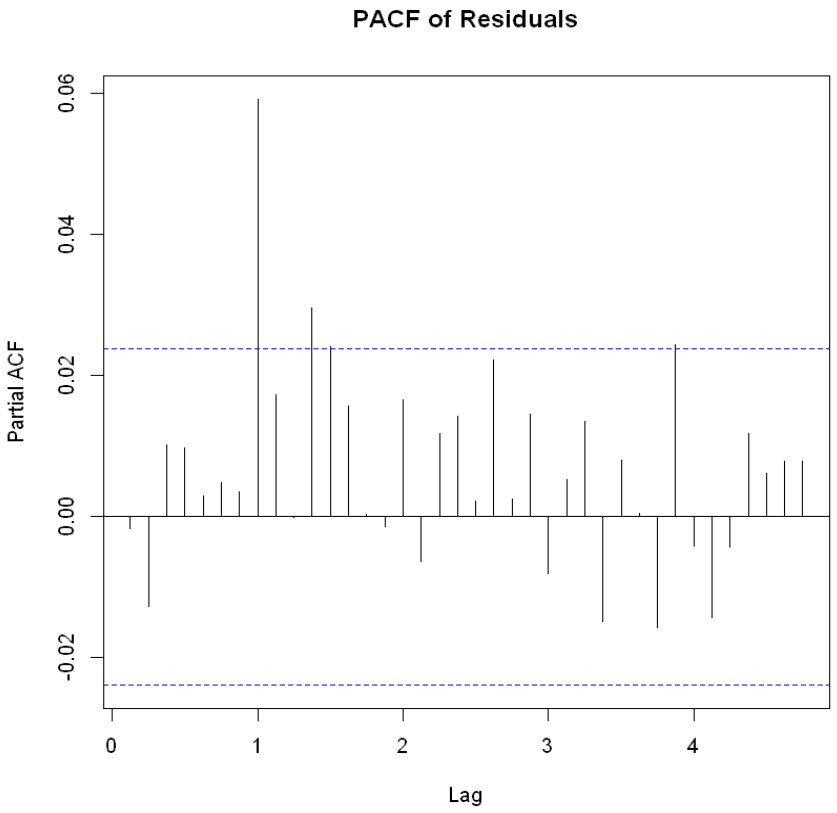


### Model 3.4. SARIMA (2,1,2)(0,1,1)[8]

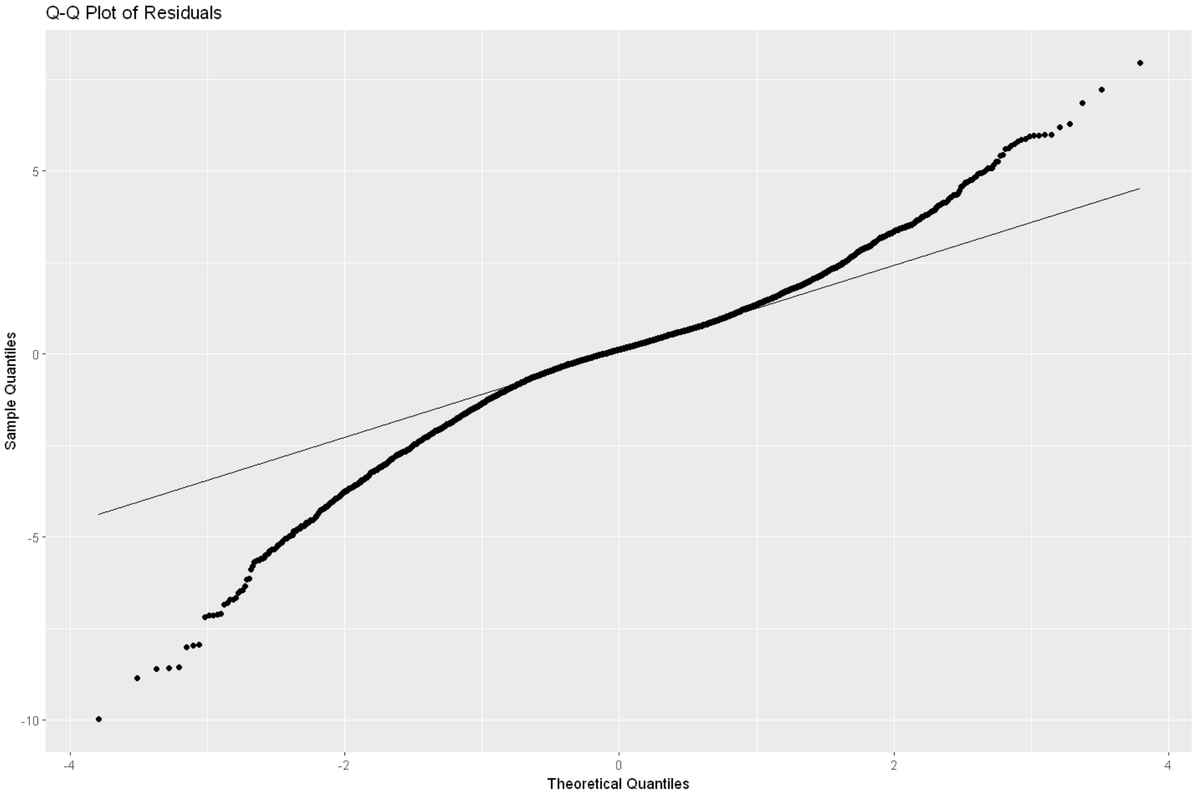
Due to the spike at PACF plot lag 2 value, AR(2) model is also integrated to obtain a more accurate model. The residual analysis and related ACF/PACF, and the Q-Q plots of each group for the finalized ARIMA model are given below.



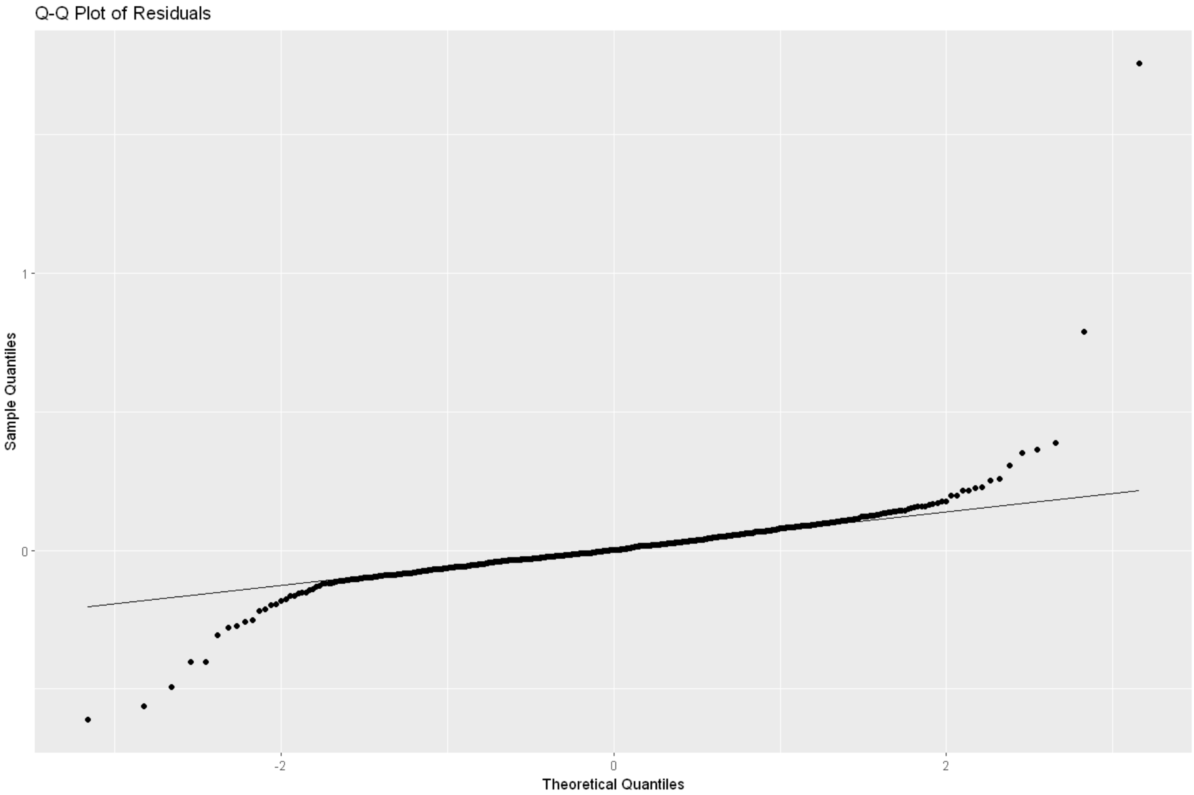

ACF of Residuals
ACF
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
Lag


PACF of Residuals
Partial ACF
0.06
0.04
0.02
0.00
-0.02
0
1
2
3
4
Lag

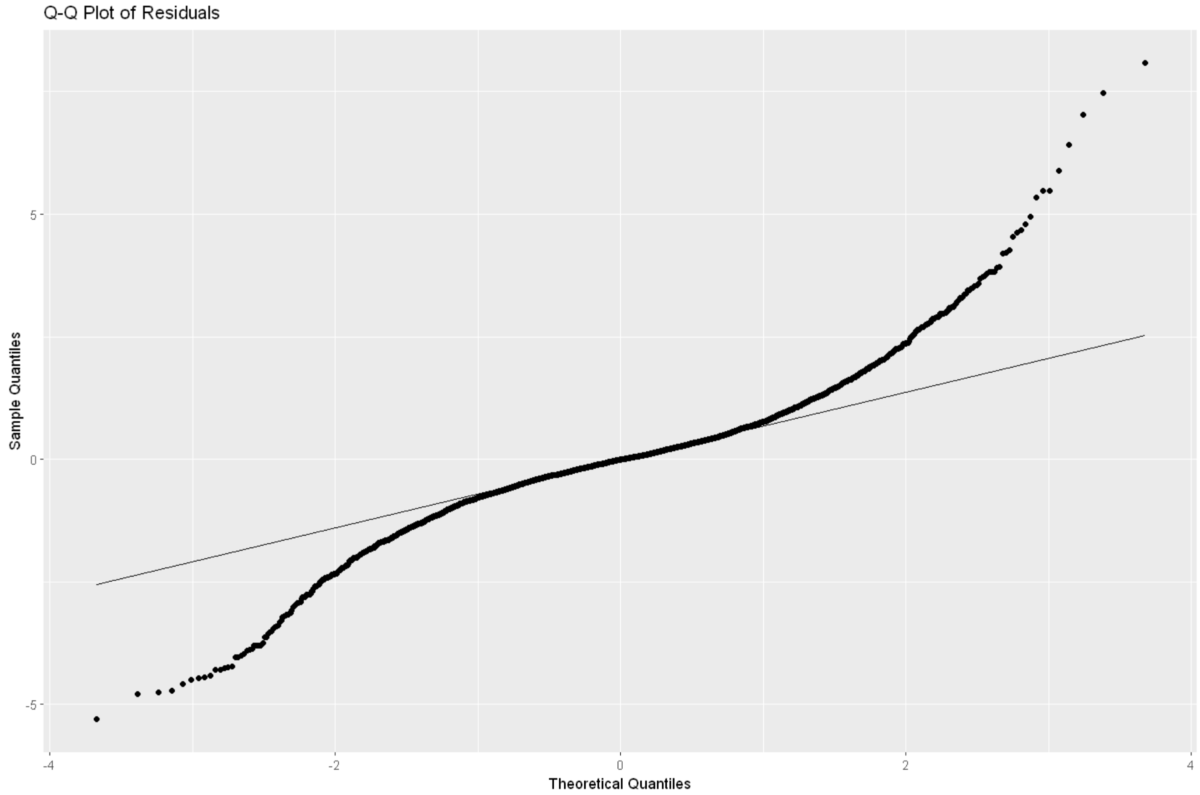
## Q-Q Plot of Group 1 (Hours 8, 9, 10, 11, 12, 13, 14, 15)



## Q-Q Plot of Group 2 (Hours 4, 5)

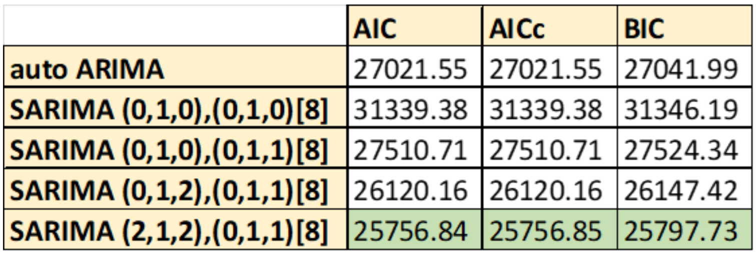


### Q-Q Plot of Group 3 (Hours 6, 7, 16, 17, 18)



## Performance Metrics Comparisons for 4 different ARIMA Models

To determine the best ARIMA model, we compared their AIC values. The AIC (Akaike Information Criterion) measures the trade-off between model fit and complexity, with lower AIC values indicating a better model that balances these aspects. By selecting the model with the lowest AIC, we aim to achieve optimal predictive performance while avoiding overfitting.



| | AIC | AICc | BIC |
|---|---|---|---|
| auto ARIMA | 27021.55 | 27021.55 | 27041.99 |
| SARIMA (0,1,0),(0,1,0)[8] | 31339.38 | 31339.38 | 31346.19 |
| SARIMA (0,1,0),(0,1,1)[8] | 27510.71 | 27510.71 | 27524.34 |
| SARIMA (0,1,2),(0,1,1)[8] | 26120.16 | 26120.16 | 26147.42 |
| SARIMA (2,1,2),(0,1,1)[8] | 25756.84 | 25756.85 | 25797.73 |

## MODEL COMPARISON

When comparing different models the following performance metrics are used; Coefficient of Variation (CV), FBias, MAPE, RMSE, MAE, MAD, MADP, and WMAPE. Each metric provides unique insights into model performance. The Coefficient of Variation (CV) standardizes the standard deviation relative to the mean, offering a measure of relative variability and allowing comparisons across datasets with different units. FBias indicates the direction and magnitude of systematic errors, revealing whether a model consistently overestimates or underestimates the true values. Mean Absolute Percentage Error (MAPE) measures the accuracy as a percentage, making it easy to interpret and compare across different contexts. Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) both quantify the average prediction error, but RMSE gives higher weight to larger errors, making it sensitive to outliers. Mean Absolute Deviation (MAD) provides a straightforward average of absolute errors, while Mean Absolute Deviation Percentage (MADP) extends this to a percentage scale for easier comparison. Weighted Mean Absolute Percentage Error (WMAPE) accounts for the importance of individual errors by weighting them, thus offering a more nuanced accuracy measure. Together, these metrics provide a comprehensive view of model performance, highlighting different aspects such as variability, bias, accuracy, and the impact of outliers. To evaluate the performances of our finalized models, all of the mentioned metric values are provided in the table below. Also, the box plots for %95 confidence interval of MAD’s and  Residual distribution graphs of each of 3 models are given.

Table of evaluation metrics of three models

Figure of Box-plots

Graph of residual distributions to be added here

## CONCLUSIONS AND FUTURE WORK

In this project, our objective was to provide accurate hourly solar power predictions for the Edikli Güneş Enerjisi Santrali (GES) located in Niğde, Turkey. We utilized historical weather data and production metrics spanning from January 1, 2022, to May 12, 2024, to develop robust forecasting models. Throughout the project, we employed three different modeling approaches: Linear Regression, ARIMA, and Regression Trees, each aimed at capturing the complex relationships between various weather factors and solar power production.

Initially, the Linear Regression model served as our baseline. Although it gave more reasonable results after integrating the dummy variables, the residual analysis and performance measure evaluations revealed that a significant seasonality still exists and there were more opportunities for decomposition.

Recognizing the limitations of our initial models, we implemented a Regression Tree model, which allowed for a more subtle understanding of the relationships between the independent variables and solar power production. By systematically analyzing the importance of each feature and refining the model accordingly, we achieved significant improvements in forecasting accuracy. However, despite these improvements, the values of the performance measures were not sufficient, indicating the need for a more sophisticated approach.

The seasonality effect was very evident in our first two models. To handle this problem, we decided to apply ARIMA method to the residuals (both with auto.arima function and deciding the parameters manually) to obtain a more proper model to handle the mentioned problems. The best results are obtained by manual decision of parameters in a step-wised way.

The SARIMA 3.4 model emerged as the most effective approach, yielding the best performance measure values. This model's success underscores the importance of considering the crucial seasonality effect on predicting solar power output. Our findings align with existing literature, which highlights the superior performance of ARIMA methods in handling complex and seasonal data patterns.

Additionally, for improving the model, one may consider using actual values from specific time periods that encompass the prediction periods. For instance, while training and evaluating a prediction for May, data from March to July could be utilized to ensure more accurate predictions. Furthermore, future work may explore integrating additional weather variables and advanced machine-learning techniques to further enhance prediction accuracy.

In conclusion, our comprehensive evaluation demonstrates that the Seasonal ARIMA model offer the best solution for solar power forecasting, outperforming both Linear Regression and Regression Tree models. The insights gained from this project contribute not only to improving the operational efficiency of Edikli GES but also to valuable knowledge to the field of sustainable energy forecasting.