It's Not Rocket Science, It's Electricity Market

Table of contents

Welcome to my EMU660 project page.

1 1. Project Overview and Scope

Electricity has increasingly become a tradable commodity on global and Turkish stock exchanges, subject to specific regulations and limitations. In a liberalized market, it is uniquely characterized by a third dimension—time—alongside price and volume. In Turkey, electricity trading takes place across multiple market platforms, all overseen and regulated by the Energy Exchange Istanbul (EXIST). This project aims to analyze the formation of electricity prices and investigate the impact of total electricity demand as well as electricity generation from various sources on price formation. Specifically, the daily impact of renewable energy generation on electricity prices will be examined, while the influence of natural gas prices on the monthly average electricity price will also be explored. Electricity prices will be forecasted using multiple linear regression models at both daily and monthly resolutions, and the results will be evaluated accordingly.

2 2. Data

This project will utilize three main data sources related to the electricity market or influence it. EPİAŞ (Energy Exchange Istanbul), TEİAŞ (Turkish Electricity Transmission Corporation), and BOTAŞ (Petroleum Pipeline Corporation) are public institutions in Turkey that act as decision-makers and regulators in the electricity market. The open-access data provided by these institutions will be used throughout the analysis and forecasting processes of the project. The whole data used in this project starts in first hour of 2023 and end at the end of 2024.

2.1 2.1 Data Source

This project will utilize three main data sources related to the electricity market or influence it. EPİAŞ (Energy Exchange Istanbul), TEİAŞ (Turkish Electricity Transmission Corporation), and BOTAŞ (Petroleum Pipeline Corporation) are public institutions in Turkey that act as decision-makers and regulators in the electricity market. The open-access data provided by these institutions will be used throughout the analysis and forecasting processes of the project. Data sources can be reached by clicking links below.

EPİAŞ

TEİAŞ

BOTAŞ

2.2 2.2 General Information About Data

Data related to the Day-Ahead Market will be obtained from EPİAŞ. The data sourced from EPİAŞ can be categorized under three main headings.

1) FDDP (Final Daily Production Program): This data is provided at an hourly resolution for 12 different types of energy sources. It includes the planned generation amounts for the following day submitted by power plants operating under each source category. Every day, power plants enter their generation schedules into the system by 4 PM, and EPİAŞ collects and publishes this data aggregated by source type.

2) Real-Time Consumption: This data represents the total amount of electricity consumed across Turkey. It is provided on an hourly basis and can be referred to as the total electricity demand.

3) MCP (Market Clearing Price): This data refers to the electricity price determined for each hour in the Day-Ahead Market, formed by matching supply and demand for the traded electricity.

Natural gas tariff data has been sourced from BOTAŞ. The prices of natural gas used for electricity generation are determined by BOTAŞ. Additionally, water

inflow data to the main basin dams, provided by TEİAŞ, may be used if deemed necessary.

2.3 2.3 Reason of Choice

The electricity market consists of various sub-markets. Making accurate price forecasts for short-term and long-term electricity sales can create significant added value. Especially in long-term purchase or sale agreements, forecasting electricity prices can facilitate more profitable commercial deals while minimizing risk. For instance, the analyses and models developed in this project can help establish a relationship between renewable energy generation and electricity prices over specific periods. These forecasts can then be used to assess buy and sell offers in the market for future periods, enabling more informed and strategic positioning.

2.4 2.4 Preprocessing

In the preprocessing stage, the data stored in Excel files was converted into RData format.

```
library(readxl)

epias_data <- read_excel("epias_data.xlsx")

botas_data <- read_excel("botas_data.xlsx")

save(epias_data, botas_data, file = "electricity.RData")

head(epias_data)
```

```
# A tibble: 6 x 18
  date                hour   total naturalgas  wind lignite darkcoal
  <dttm>              <chr>  <dbl>      <dbl> <dbl>   <dbl>    <dbl>
1 2023-01-01 00:00:00 00:00 25945.      3636.  988.   4646.      136
2 2023-01-01 01:00:00 01:00 24494.      2939. 1056.   4646.      136
3 2023-01-01 02:00:00 02:00 22631.      2495. 1123.   4694.      136
4 2023-01-01 03:00:00 03:00 22022.      2667. 1227.   4878.      136
5 2023-01-01 04:00:00 04:00 21404.      2334. 1308.   4910.      136
6 2023-01-01 05:00:00 05:00 21586.      2177. 1406.   5016.      136
# i 11 more variables: importedcoal <dbl>, fueloil <dbl>, geothermal <dbl>,
#   dam <dbl>, naphta <dbl>, biomass <dbl>, runofriver <dbl>, other <dbl>,
#   demand <dbl>, solar <dbl>, price <dbl>
```

```
head(botas_data)
```

```
# A tibble: 6 x 3
   year month natgasprice
```

```
  <dbl> <dbl>       <dbl>
1  2023     1       18000
2  2023     2       15000
3  2023     3       12000
4  2023     4       10000
5  2023     5       10000
6  2023     6       10000
```

The natural gas price data obtained from BOTAŞ was available on a monthly basis. These monthly values were integrated into the hourly dataset.

```
load("electricity.RData")
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
epias_data$date <- as.Date(epias_data$date)

epias_data$year <- format(epias_data$date, "%Y")
epias_data$month <- format(epias_data$date, "%m")

botas_data$month <- sprintf("%02d", botas_data$month)

botas_data$year <- as.character(botas_data$year)

epias_merged <- left_join(epias_data, botas_data, by = c("year", "month"))

head(epias_merged)
```

```
# A tibble: 6 x 21
  date       hour   total naturalgas  wind lignite darkcoal importedcoal fueloil
  <date>     <chr>  <dbl>      <dbl> <dbl>   <dbl>    <dbl>        <dbl>   <dbl>
1 2023-01-01 00:00 25945.      3636.  988.   4646.      136        8556.    16.5
2 2023-01-01 01:00 24494.      2939. 1056.   4646.      136        8556.    16.5
3 2023-01-01 02:00 22631.      2495. 1123.   4694.      136        8554.    16.5
4 2023-01-01 03:00 22022.      2667. 1227.   4878.      136        8372.    16.5
5 2023-01-01 04:00 21404.      2334. 1308.   4910.      136        8373.    16.5
6 2023-01-01 05:00 21586.      2177. 1406.   5016.      136        8373.    15.5
# i 12 more variables: geothermal <dbl>, dam <dbl>, naphta <dbl>,
```

```
#   biomass <dbl>, runofriver <dbl>, other <dbl>, demand <dbl>, solar <dbl>,
#   price <dbl>, year <chr>, month <chr>, natgasprice <dbl>
```

```
save(epias_merged, file = "electricity_merged.RData")

epias_merged <- epias_merged %>%
  select(date, year, month, everything(), -year, -month)

epias_merged <- epias_merged %>%
  select(-price, everything(), price)

save(epias_merged, file = "electricity_merged.RData")
```

The hourly data was converted into daily averages to enable analysis at a daily resolution.

```
library(dplyr)

epias_daily <- epias_merged %>%
  group_by(date) %>%
  summarise(across(where(is.numeric), mean, na.rm = TRUE))
```

```
Warning: There was 1 warning in `summarise()`.
i In argument: `across(where(is.numeric), mean, na.rm = TRUE)`.
i In group 1: `date = 2023-01-01`.
Caused by warning:
! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.
Supply arguments directly to `.fns` through an anonymous function instead.

  # Previously
  across(a:b, mean, na.rm = TRUE)

  # Now
  across(a:b, \(x) mean(x, na.rm = TRUE))
```

Finally, a feature aggregation process was carried out to prepare the data for forecasting. Generation sources with similar characteristics were grouped under common categories. Fueloil, naphta, lignite, and hard coal were combined under the label cheap_thermal. Wind, run-of-river, biomass, and geothermal sources were grouped under renewables.

Solar and hydro (dam) generation were excluded from the renewables group, as they exhibit distinct production characteristics. The category other was disregarded due to its low share and lack of detailed classification.

```
library(dplyr)

epias_simplified_daily <- epias_daily %>%
```

```
  mutate(
    cheap_thermal = fueloil + naphta + lignite + darkcoal,
    renewables = wind + runofriver + biomass + geothermal
  ) %>%
  select(
    date, cheap_thermal, renewables,
    importedcoal, naturalgas, solar, dam, demand, natgasprice, price
  )
save(epias_simplified_daily, file = "epias_simplified_daily.RData")

head(epias_simplified_daily)
```

```
# A tibble: 6 x 10
  date       cheap_thermal renewables importedcoal naturalgas solar   dam demand
  <date>             <dbl>      <dbl>        <dbl>      <dbl> <dbl> <dbl>  <dbl>
1 2023-01-01         4972.      4375.        8332.      3210. 1491. 3100. 28743.
2 2023-01-02         5128.      3919.        9062.      8888. 1441. 2541. 35772.
3 2023-01-03         5016.      4079.        9328.     11235. 1417. 2531. 37497.
4 2023-01-04         4834.      5007.        8810.     10960. 1345. 2213. 38064.
5 2023-01-05         4849.      4986.        8435.     11628. 1207. 2367. 37877.
6 2023-01-06         4950.      4648.        8633.     11608. 1126. 2334. 37821.
# i 2 more variables: natgasprice <dbl>, price <dbl>
```

3 3. Analysis

3.1 3.1 Why It's Important to Create a Forecast?

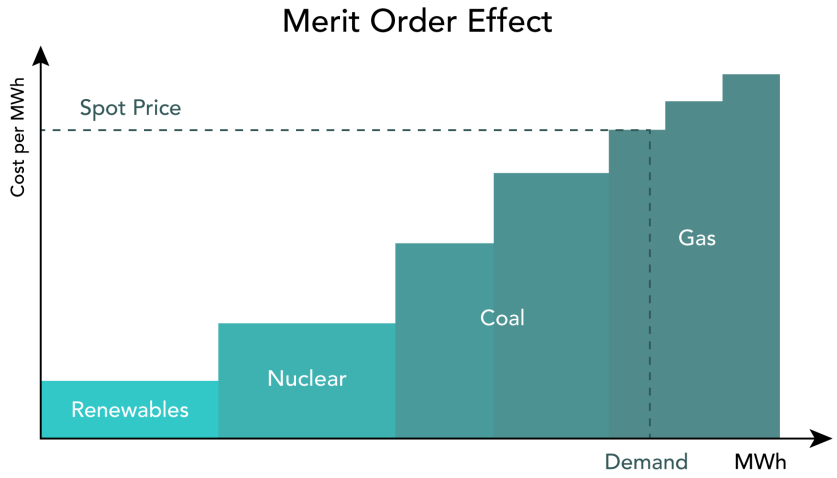
Electricity price forecasting plays a critical role in the power market. Accurate forecasts are essential not only for making commercial decisions but also for managing financial processes. Cash flow management is a key factor for maintaining an active and balanced presence in the market. Portfolios with constant inflows and outflows must manage their commercial balance while simultaneously overseeing their cash flows. Daily electricity price forecasting can provide significant advantages in both commercial and financial foresight.

3.2 3.2 How Electricity Prices are Determined?

Electricity prices are determined through a system known as the merit order. Similar to basic economic pricing, supply and demand are matched for each hour, and the price is set at the point where supply meets demand. However, unlike other markets, electricity is a fundamental need—so demand does not respond to price, but instead determines it.

For any given hour, demand is met starting from the cheapest suppliers, moving up to the more expensive ones. The price of electricity is then set based on the bid of the most expensive accepted supplier. Electricity generation resources

can generally be ranked from cheapest to most expensive as follows: renewables, hydro, nuclear, domestic coal, imported coal, and natural gas.



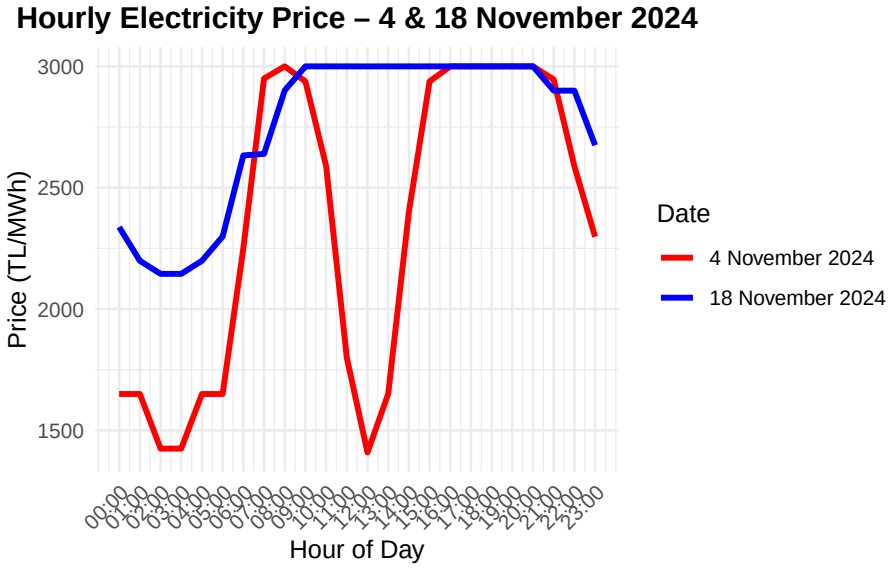
3.3 3.3 Exploratory Data Analysis & Trend Analysis

Electricity prices can be highly volatile even on the same type of day. The chart below shows the prices for two different Mondays in November 2024

```
library(dplyr)
library(ggplot2)

# Veri seçimi
price_data <- epias_merged %>%
  filter(date %in% as.Date(c("2024-11-04", "2024-11-18"))) %>%
  mutate(hour_numeric = as.numeric(substr(hour, 1, 2))) %>%
  select(date, hour_numeric, price)

# Grafik
ggplot(price_data, aes(x = hour_numeric, y = price, color = as.factor(date), group = date))
  geom_line(size = 1.2) +
  scale_color_manual(
    values = c("2024-11-04" = "red", "2024-11-18" = "blue"),
    labels = c("4 November 2024", "18 November 2024")
  ) +
  scale_x_continuous(
    breaks = 0:23,
    labels = sprintf("%02d:00", 0:23)
  ) +
  labs(
```

```
    title = "Hourly Electricity Price - 4 & 18 November 2024",
    x = "Hour of Day",
    y = "Price (TL/MWh)",
    color = "Date"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



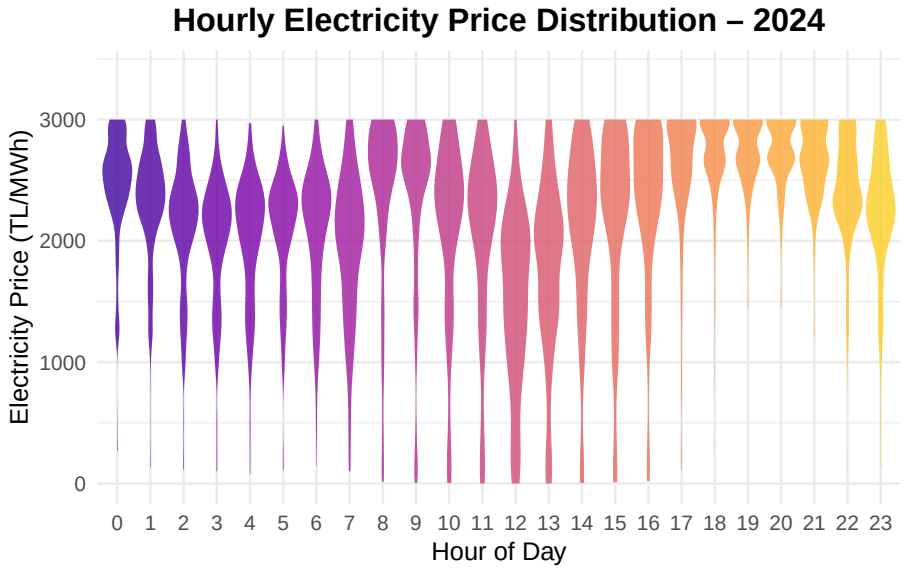
Electricity prices exhibit high volatility across different hours of the day. This volatility is clearly visible in the chart below, with midday hours standing out as particularly volatile.

```
library(dplyr)
library(ggplot2)
library(lubridate)
```

```
Attaching package: 'lubridate'
```

```
The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

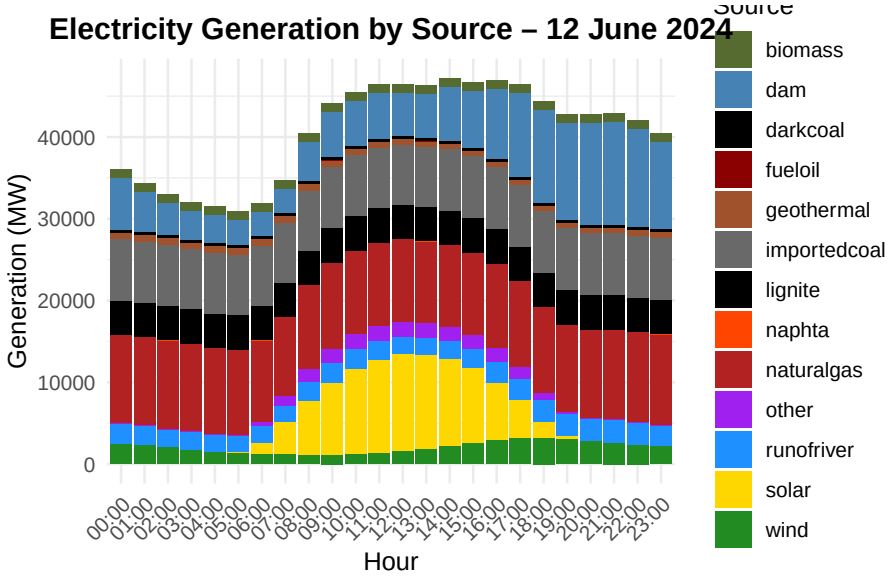
```r
# 1. Saat sütunu sayıya çevrilir
epias_merged <- epias_merged %>%
  mutate(hour_numeric = as.numeric(substr(hour, 1, 2)))

# 2. Sadece 2024 yılı alınır
price_2024 <- epias_merged %>%
  filter(year(date) == 2024) %>%
  select(hour_numeric, price)

# 3. Violin plot: her saat için fiyat yoğunluğu
ggplot(price_2024, aes(x = factor(hour_numeric), y = price, fill = factor(hour_numeric))) +
  geom_violin(scale = "width", adjust = 1.2, alpha = 0.8, color = NA) +
  scale_fill_viridis_d(option = "C", begin = 0.1, end = 0.9) +
  coord_cartesian(ylim = c(0, 3400)) +  # Görsel netlik için üst sınır
  labs(
    title = "Hourly Electricity Price Distribution - 2024",
    x = "Hour of Day",
    y = "Electricity Price (TL/MWh)",
    fill = "Hour"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
    plot.subtitle = element_text(hjust = 0.5),
    legend.position = "none"
  )
```



To understand the electricity market, it is essential to first examine the factors that influence it. Among these, the most critical elements are generation and consumption data.

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(lubridate)

prod_data <- epias_merged %>%
  filter(date == as.Date("2024-06-12")) %>%
  select(hour, solar, wind, runofriver, dam, geothermal, biomass, naturalgas,
         fueloil, naphta, lignite, darkcoal, importedcoal, other)

prod_long <- prod_data %>%
  pivot_longer(
    cols = -hour,
    names_to = "source",
    values_to = "generation"
  )

source_colors <- c(
  solar = "gold",
  wind = "forestgreen",
  runofriver = "dodgerblue",
  dam = "steelblue",
  geothermal = "sienna",
  biomass = "darkolivegreen",
  naturalgas = "firebrick",
  fueloil = "darkred",
  naphta = "orangered",
  lignite = "black",
  darkcoal = "black",
  importedcoal = "dimgray",
  other = "purple"
)

ggplot(prod_long, aes(x = hour, y = generation, fill = source)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = source_colors) +
  labs(
    title = "Electricity Generation by Source - 12 June 2024",
    x = "Hour",
    y = "Generation (MW)",
    fill = "Source"
  ) +
```

```
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```



This chart clearly shows how electricity demand is met by different generation sources at different hours of the day.

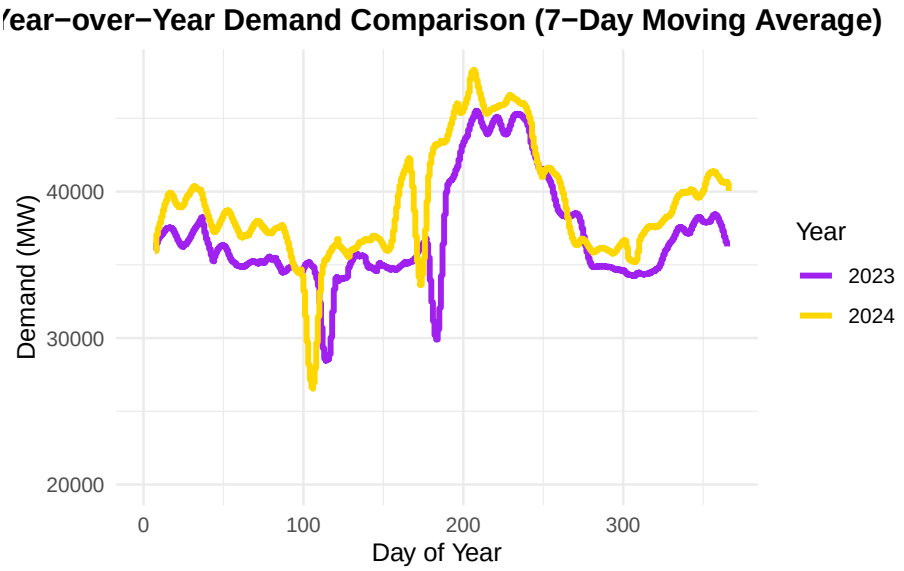
There are two major factors that influence electricity consumption, or demand. The first is air temperature. In Türkiye, electricity demand increases when the average daily temperature rises above or drops below 15°C, due to higher use of heating and cooling systems.

The second factor is more long-term: the country's level of industrial activity and population size. As the population grows and industrial production expands, electricity demand also rises.

The chart below displays the moving average electricity demand data for 2023 and 2024.

```
library(dplyr)
library(ggplot2)
library(lubridate)
library(zoo)

demand_yoy <- epias_merged %>%
  filter(year(date) %in% c(2023, 2024)) %>%
```

```
  arrange(date, hour) %>%
  mutate(
    year = year(date),
    doy = yday(date),
    datetime = as.POSIXct(paste(date, hour), format = "%Y-%m-%d %H:%M")
  ) %>%
  group_by(year) %>%
  arrange(datetime) %>%
  mutate(
    demand_7day_avg = rollmean(demand, k = 24 * 7, fill = NA, align = "right")
  ) %>%
  ungroup()

ggplot(demand_yoy, aes(x = doy, y = demand_7day_avg, color = as.factor(year))) +
  geom_line(size = 1.2) +
  labs(
    title = "Year-over-Year Demand Comparison (7-Day Moving Average)",
    x = "Day of Year",
    y = "Demand (MW)",
    color = "Year"
  ) +
  scale_color_manual(values = c("2023" = "purple", "2024" = "gold")) +
  expand_limits(y = 20000) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold")
  )
```

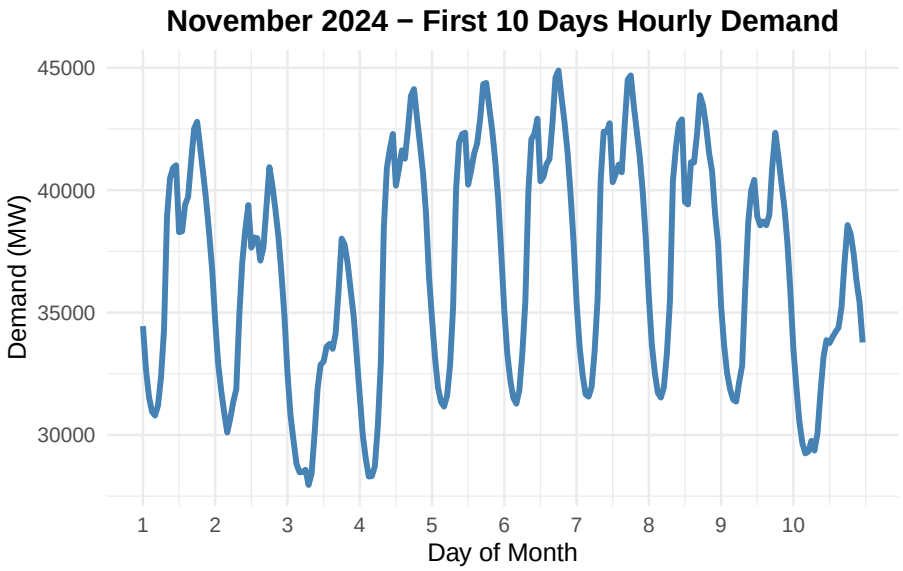


As seen in the chart, electricity demand increases during the summer due to rising temperatures, and also in winter as temperatures drop. Additionally, demand tends to reach its lowest levels during religious holidays, when industrial activity comes to a near halt.

When examining demand data at a higher resolution, it becomes clear how electricity demand varies across the days of the week.

```
library(dplyr)
library(ggplot2)
library(lubridate)

nov_demand <- epias_merged %>%
  filter(date >= as.Date("2024-11-01") & date <= as.Date("2024-11-10")) %>%
  mutate(
    datetime = as.POSIXct(paste(date, hour), format = "%Y-%m-%d %H:%M"),
    day_num = day(date) + hour(datetime) / 24
  )

ggplot(nov_demand, aes(x = day_num, y = demand)) +
  geom_line(color = "steelblue", size = 1.2) +
  scale_x_continuous(breaks = 1:10) +
  labs(
    title = "November 2024 - First 10 Days Hourly Demand",
    x = "Day of Month",
    y = "Demand (MW)"
  ) +
```

```
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold")
  )
```



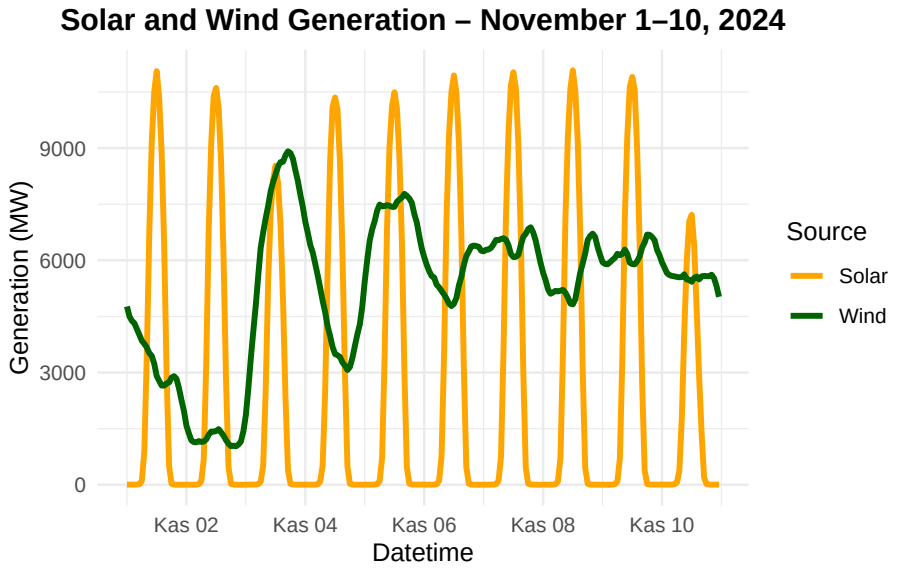
Electricity demand remains relatively consistent during weekdays, whereas a noticeable drop is observed on Saturdays and Sundays.

A comprehensive understanding of the data requires analyzing generation figures in conjunction with demand.

Renewable energy sources exhibit different generation trend characteristics. Among these, the most influential factors are meteorological conditions and seasonal effects. Below is the generation data from various sources for a single day. As seen, solar production peaks around midday and drops to zero after sunset. Wind generation, on the other hand, may display varying patterns from day to day.

```
library(dplyr)
library(ggplot2)
library(lubridate)

nov_renew <- epias_merged %>%
  filter(date >= as.Date("2024-11-01") & date <= as.Date("2024-11-10")) %>%
  mutate(
    datetime = as.POSIXct(paste(date, hour), format = "%Y-%m-%d %H:%M")
  )
```

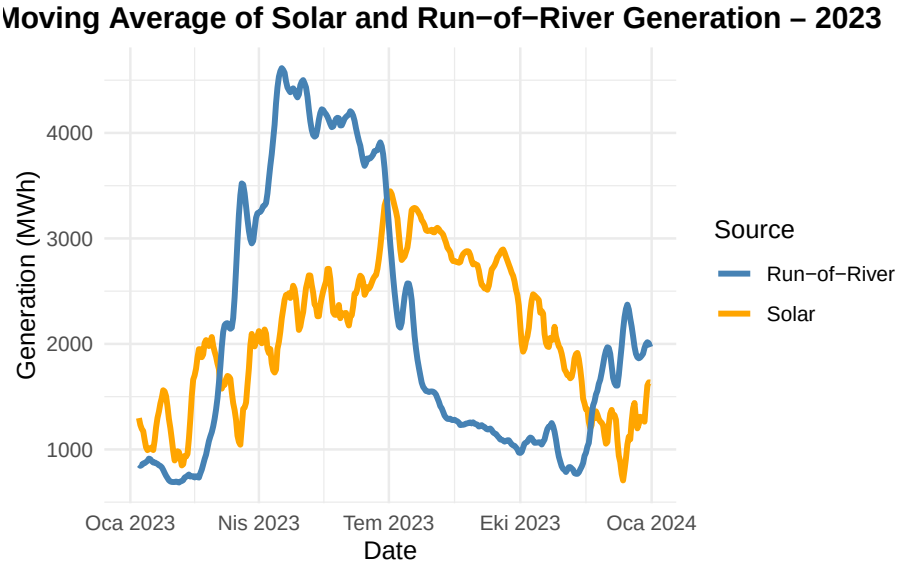
```
ggplot(nov_renew, aes(x = datetime)) +
  geom_line(aes(y = solar, color = "Solar"), size = 1.2) +
  geom_line(aes(y = wind, color = "Wind"), size = 1.2) +
  scale_color_manual(values = c("Solar" = "orange", "Wind" = "darkgreen")) +
  labs(
    title = "Solar and Wind Generation - November 1-10, 2024",
    x = "Datetime",
    y = "Generation (MW)",
    color = "Source"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold")
  )
```



An analysis of solar and run-of-river generation data reveals clear seasonal trends. While solar production increases during the summer months, run-of-river generation peaks in the spring.

```
library(dplyr)
library(ggplot2)
library(zoo)
library(lubridate)

solar_river_daily <- epias_merged %>%
  mutate(date = as.Date(date)) %>%
  group_by(date) %>%
```

```r
  summarise(
    solar = mean(solar, na.rm = TRUE),
    runofriver = mean(runofriver, na.rm = TRUE)
  ) %>%
  filter(year(date) == 2023) %>%
  arrange(date) %>%
  mutate(
    solar_ma = rollmean(solar, k = 7, fill = NA, align = "right"),
    runofriver_ma = rollmean(runofriver, k = 7, fill = NA, align = "right")
  )

ggplot(solar_river_daily, aes(x = date)) +
  geom_line(aes(y = solar_ma, color = "Solar"), size = 1.2) +
  geom_line(aes(y = runofriver_ma, color = "Run-of-River"), size = 1.2) +
  scale_color_manual(values = c("Solar" = "orange", "Run-of-River" = "steelblue")) +
  labs(
    title = "7-Day Moving Average of Solar and Run-of-River Generation - 2023",
    x = "Date",
    y = "Generation (MWh)",
    color = "Source"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



3.4 3.3 Model Fitting

Daily average electricity prices were forecasted using a multiple linear regression approach.

Two different models were developed. Although both models followed the same computational logic, they differed in feature selection and feature aggregation strategies. In the first model, all available variables were used individually as features, whereas in the second model, a feature aggregation approach was applied.

3.3.1 What Happens If We Don't Aggregate Features?

```
library(dplyr)
library(car)
```

```
Zorunlu paket yükleniyor: carData
```

```

Attaching package: 'car'
```

```
The following object is masked from 'package:dplyr':

    recode
```

```
library(ggplot2)

reg_data1 <- epias_daily %>%
  select(price, naturalgas, wind, lignite, darkcoal, importedcoal, fueloil, geothermal, dam,
  na.omit()

model1 <- lm(price ~ naturalgas + wind + lignite + darkcoal + importedcoal + fueloil + geoth

summary(model1)
```

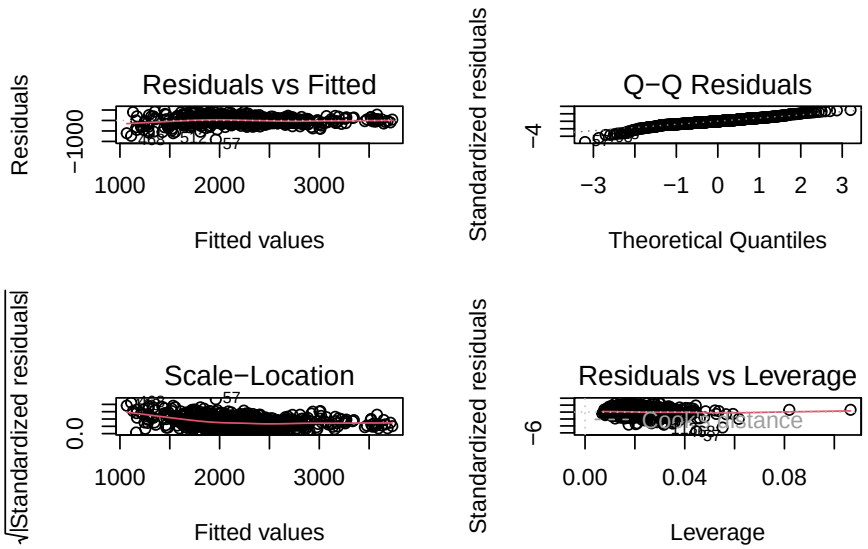
```

Call:
lm(formula = price ~ naturalgas + wind + lignite + darkcoal +
    importedcoal + fueloil + geothermal + dam + naphta + biomass +
    runofriver + other + demand + solar + natgasprice, data = reg_data1)

Residuals:
    Min      1Q  Median      3Q     Max
-913.47  -89.95   -1.53  101.26  501.77

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.982e+03  2.039e+02  -9.725  < 2e-16 ***
naturalgas    4.853e-02  1.083e-02   4.481 8.63e-06 ***
```

```
wind         -4.299e-02  1.021e-02  -4.212 2.85e-05 ***
lignite       1.250e-02  2.292e-02   0.545  0.58566
darkcoal      1.293e-01  9.810e-02   1.318  0.18779
importedcoal  7.100e-02  1.422e-02   4.993 7.47e-07 ***
fueloil       2.881e+00  1.219e+00   2.364  0.01834 *
geothermal    2.114e-01  1.384e-01   1.528  0.12707
dam          -5.017e-02  1.136e-02  -4.415 1.16e-05 ***
naphta       -4.473e+01  6.276e+01  -0.713  0.47623
biomass       8.360e-04  1.765e-01   0.005  0.99622
runofriver    1.168e-01  1.824e-02   6.404 2.74e-10 ***
other        -4.118e-02  8.859e-02  -0.465  0.64217
demand        3.037e-02  9.991e-03   3.040  0.00245 **
solar         2.519e-02  1.683e-02   1.496  0.13501
natgasprice   1.701e-01  6.133e-03  27.736  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 168.9 on 715 degrees of freedom
Multiple R-squared:  0.8836,    Adjusted R-squared:  0.8812
F-statistic: 361.9 on 15 and 715 DF,  p-value: < 2.2e-16
```

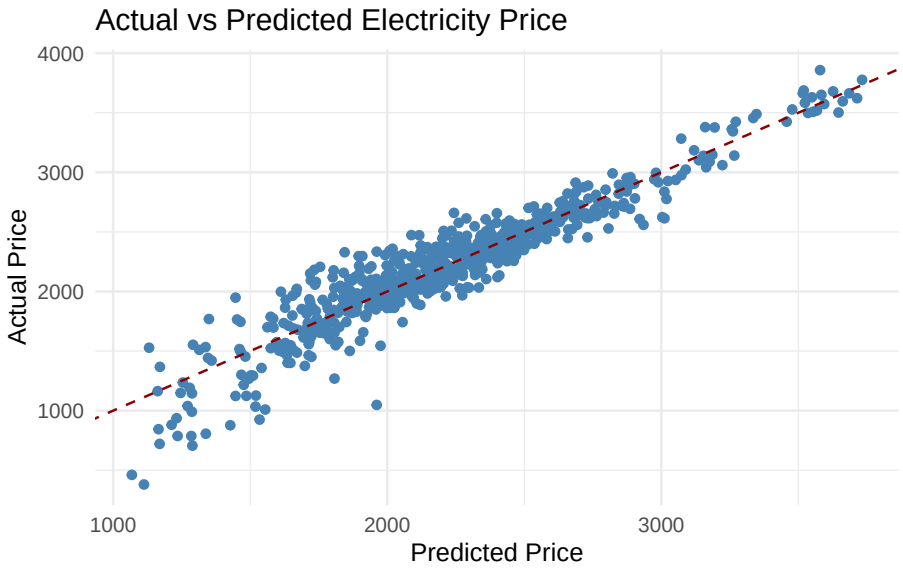
```r
vif(model1)
```

```
  naturalgas         wind      lignite     darkcoal importedcoal      fueloil
   36.019212    11.420398     3.354866     1.503104    13.072816     1.279864
  geothermal          dam       naphta      biomass   runofriver        other
    5.939204    10.702876     1.057892     2.303100    11.131208     4.756550
      demand        solar  natgasprice
   51.610155     8.003321     2.753763
```

```r
par(mfrow = c(2, 2))
plot(model1)
```




```r
par(mfrow = c(1, 1))


reg_data1$predicted_price1 <- predict(model1)

ggplot(reg_data1, aes(x = predicted_price1, y = price)) +
  geom_point(color = "steelblue") +
  geom_abline(slope = 1, intercept = 0, color = "darkred", linetype = "dashed") +
  labs(
    title = "Actual vs Predicted Electricity Price",
    x = "Predicted Price",
    y = "Actual Price"
  ) +
  theme_minimal()
```



3.3.2 What If We Aggregate Features?

```
library(dplyr)
library(car)
library(ggplot2)

reg_data <- epias_simplified_daily %>%
  select(price, cheap_thermal, naturalgas, importedcoal, renewables, demand, solar, dam, nat
  na.omit()


model <- lm(price ~ cheap_thermal + naturalgas + importedcoal + renewables + demand + solar

summary(model)
```

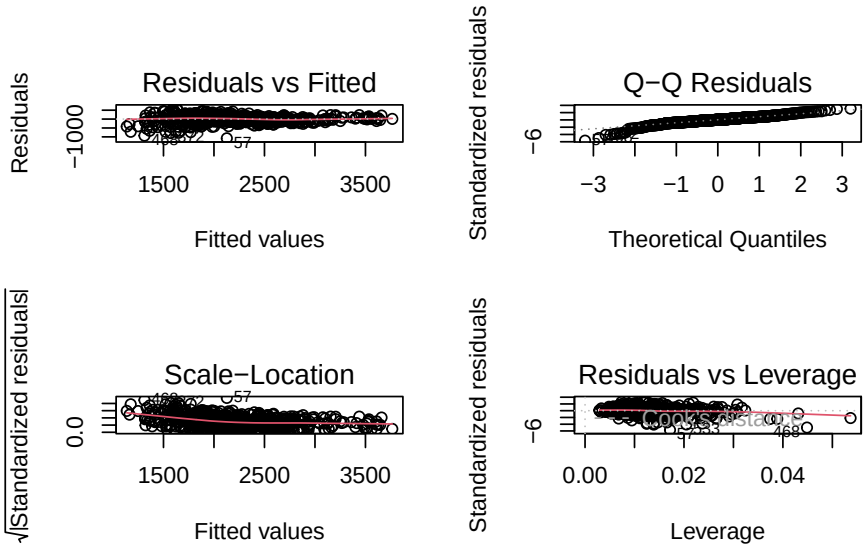
```
Call:
lm(formula = price ~ cheap_thermal + naturalgas + importedcoal +
    renewables + demand + solar + dam + natgasprice, data = reg_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1078.97   -97.87     7.74   114.92   587.23

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.994e+02  1.027e+02  -6.812 2.03e-11 ***
```

```
cheap_thermal  5.054e-03  2.295e-02   0.220 0.825754
naturalgas     1.977e-02  1.026e-02   1.926 0.054483 .
importedcoal  -3.961e-02  1.080e-02  -3.666 0.000264 ***
renewables    -7.091e-02  9.712e-03  -7.302 7.52e-13 ***
demand         5.646e-02  8.919e-03   6.330 4.29e-10 ***
solar         -4.154e-02  1.038e-02  -4.003 6.91e-05 ***
dam           -4.511e-02  1.023e-02  -4.412 1.18e-05 ***
natgasprice    1.544e-01  4.778e-03  32.311  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 191.1 on 722 degrees of freedom
Multiple R-squared:  0.8494,    Adjusted R-squared:  0.8477
F-statistic: 509.1 on 8 and 722 DF,  p-value: < 2.2e-16
```

```r
vif(model)
```

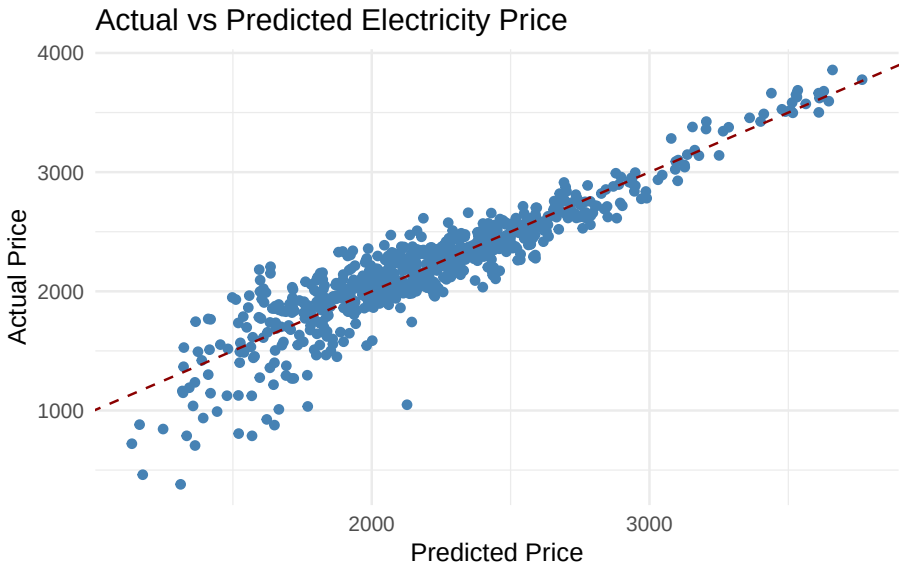
```
cheap_thermal     naturalgas   importedcoal     renewables         demand
     2.984234      25.250485       5.887888       9.567100      32.098596
        solar            dam   natgasprice
     2.374800       6.767137       1.304383
```

```r
par(mfrow = c(2, 2))
plot(model)
```




```r
par(mfrow = c(1, 1))
```

```r
reg_data$predicted_price <- predict(model)

ggplot(reg_data, aes(x = predicted_price, y = price)) +
  geom_point(color = "steelblue") +
  geom_abline(slope = 1, intercept = 0, color = "darkred", linetype = "dashed") +
  labs(
    title = "Actual vs Predicted Electricity Price",
    x = "Predicted Price",
    y = "Actual Price"
  ) +
  theme_minimal()
```



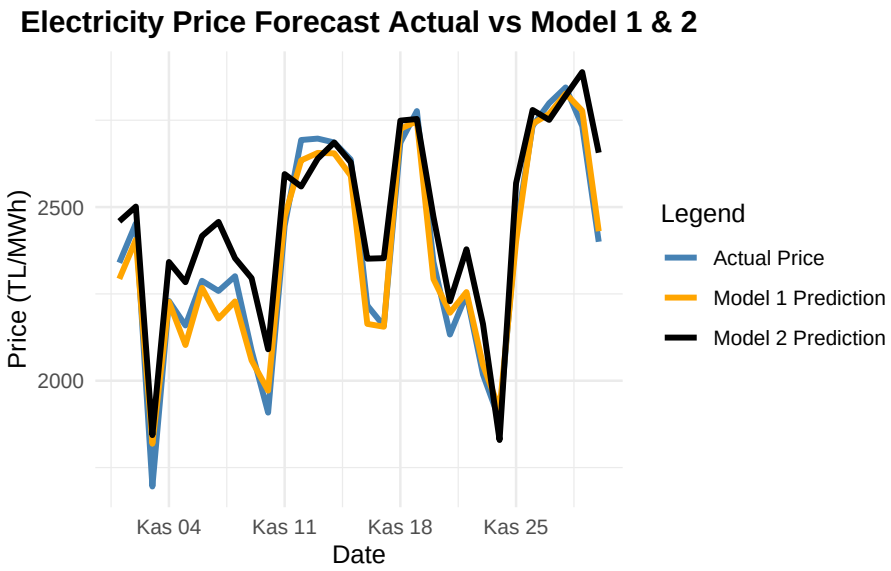
3.5 3.4 Results

When both models are evaluated, the model without feature aggregation shows a higher R-squared value and a lower standard error, indicating better predictive performance. However, an inspection of the feature coefficients reveals that the first model includes more features with statistically insignificant coefficients. Furthermore, the near-zero effect of solar generation in this model raises questions about its reliability.

Although the second model exhibits a slightly higher error rate, it offers a clearer explanation of how each feature impacts the electricity price and stands out for its simpler structure. Despite this, both models appear to produce somewhat biased predictions for extreme price values (very high or very low), likely due to the influence of outliers.

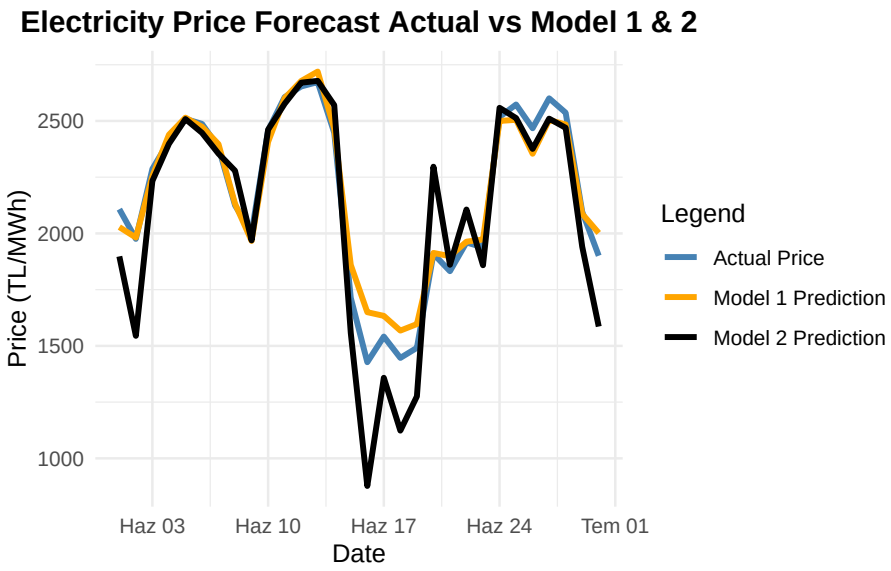
The chart below presents the actual and predicted electricity prices for November and June 2024. While both models demonstrate a strong ability to track the general trend, Model 2 shows a greater tendency toward producing outlier or extreme values.

```
data1 <- epias_daily %>%
  select(date, price, naturalgas, wind, lignite, darkcoal, importedcoal, fueloil, geothermal
  na.omit() %>%
  mutate(predicted_price1 = predict(model1, newdata = .))

data2 <- epias_simplified_daily %>%
  select(date, price, cheap_thermal, naturalgas, importedcoal, renewables, demand, solar, da
  na.omit() %>%
  mutate(predicted_price2 = predict(model, newdata = .))


combined <- data1 %>%
  inner_join(data2 %>% select(date, predicted_price2), by = "date") %>%
  filter(date >= as.Date("2024-11-01") & date <= as.Date("2024-11-30"))


library(tidyr)
plot_data <- combined %>%
  select(date, price, predicted_price1, predicted_price2) %>%
  pivot_longer(cols = c(price, predicted_price1, predicted_price2),
               names_to = "type", values_to = "value")


ggplot(plot_data, aes(x = date, y = value, color = type)) +
  geom_line(size = 1.2) +
  scale_color_manual(
    values = c("price" = "black", "predicted_price1" = "steelblue", "predicted_price2" = "or
    labels = c("Actual Price", "Model 1 Prediction", "Model 2 Prediction")
  ) +
  labs(
    title = "Electricity Price Forecast Actual vs Model 1 & 2",
    x = "Date",
    y = "Price (TL/MWh)",
    color = "Legend"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```




```
data1 <- epias_daily %>%
  select(date, price, naturalgas, wind, lignite, darkcoal, importedcoal, fueloil, geothermal
  na.omit() %>%
  mutate(predicted_price1 = predict(model1, newdata = .))

data2 <- epias_simplified_daily %>%
  select(date, price, cheap_thermal, naturalgas, importedcoal, renewables, demand, solar, da
  na.omit() %>%
  mutate(predicted_price2 = predict(model, newdata = .))


combined <- data1 %>%
  inner_join(data2 %>% select(date, predicted_price2), by = "date") %>%
  filter(date >= as.Date("2024-06-01") & date <= as.Date("2024-06-30"))


library(tidyr)
plot_data <- combined %>%
  select(date, price, predicted_price1, predicted_price2) %>%
  pivot_longer(cols = c(price, predicted_price1, predicted_price2),
               names_to = "type", values_to = "value")


ggplot(plot_data, aes(x = date, y = value, color = type)) +
  geom_line(size = 1.2) +
  scale_color_manual(
```

```
    values = c("price" = "black", "predicted_price1" = "steelblue", "predicted_price2" = "or
    labels = c("Actual Price", "Model 1 Prediction", "Model 2 Prediction")
  ) +
  labs(
    title = "Electricity Price Forecast Actual vs Model 1 & 2",
    x = "Date",
    y = "Price (TL/MWh)",
    color = "Legend"
  ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```




```
library(dplyr)
library(lubridate)
library(knitr)
library(kableExtra)
```

```
Attaching package: 'kableExtra'
```

```
The following object is masked from 'package:dplyr':

    group_rows
```

```
# MAPE fonksiyonu
mape <- function(actual, predicted) {
  mean(abs((actual - predicted) / actual), na.rm = TRUE) * 100
}
```

```r
# Tahminlerin eklendiği ve birleştirilmiş veri
combined_data <- epias_daily %>%
  select(date, price,
         naturalgas, wind, lignite, darkcoal, importedcoal, fueloil,
         geothermal, dam, naphta, biomass, runofriver, other, demand, solar, natgasprice) %>
  na.omit() %>%
  mutate(pred_model1 = predict(model1, newdata = .)) %>%
  inner_join(
    epias_simplified_daily %>%
      select(date,
             cheap_thermal, naturalgas, importedcoal, renewables, demand, solar, dam, natgas
      na.omit() %>%
      mutate(pred_model2 = predict(model, newdata = .)),
    by = "date"
  )

# Aylık MAPE hesaplama
mape_summary <- combined_data %>%
  mutate(month = floor_date(date, "month")) %>%
  group_by(month) %>%
  summarise(
    MAPE_Model1 = mape(price, pred_model1),
    MAPE_Model2 = mape(price, pred_model2)
  )

# Ortalama MAPE satırını ekle
mape_summary_final <- bind_rows(
  mape_summary,
  summarise(mape_summary,
            month = as.Date("9999-12-31"),
            MAPE_Model1 = mean(MAPE_Model1),
            MAPE_Model2 = mean(MAPE_Model2))
)

# Tarih sütununu karakter formatına çevir
mape_summary_final <- mape_summary_final %>%
  mutate(month = if_else(month == as.Date("9999-12-31"), "Average", format(month, "%Y-%m")))

# Tabloyu yazdır
kable(mape_summary_final, digits = 2, caption = "Monthly MAPE Comparison of Model 1 and Mode
  kable_styling(full_width = FALSE, position = "center") %>%
  row_spec(nrow(mape_summary_final), bold = TRUE, background = "#f2f2f2",
           extra_css = "border-top: 2px solid #999;")
```

Table 1: Monthly MAPE Comparison of Model 1 and Model 2

month	MAPE_Model1	MAPE_Model2
2023-01	3.11	2.71
2023-02	6.06	6.66
2023-03	7.05	5.51
2023-04	17.37	21.19
2023-05	9.31	11.19
2023-06	9.52	11.38
2023-07	6.01	7.88
2023-08	4.85	6.78
2023-09	4.64	4.53
2023-10	4.29	3.17
2023-11	6.27	7.84
2023-12	3.98	6.02
2024-01	6.74	9.20
2024-02	8.44	9.70
2024-03	5.82	5.49
2024-04	17.65	22.54
2024-05	8.61	9.92
2024-06	8.79	10.63
2024-07	3.29	3.51
2024-08	3.06	3.86
2024-09	3.89	3.94
2024-10	6.23	7.30
2024-11	4.62	4.58
2024-12	5.41	4.08
Average	**6.88**	**7.90**

MAPE, the most widely used performance metric in electricity price forecasting, reveals an approximate 1% difference in error between the two models. The model without feature aggregation outperforms the aggregated one, achieving the lowest error rate at 6.88%.

In both of models, Natural Gas Price is the key indicator of prediction.

4 4. Results and Key Takeaways

This project focused on forecasting electricity prices in Turkey's day-ahead market. The main objective was to demonstrate that accurate price forecasting in this market does not necessarily require highly complex models, and that successful results can be achieved with relatively simple approaches. A forecasting

model was developed using publicly available data from EPİAŞ and BOTAŞ. Two different models were built, with the only distinction being their approach to feature usage. While one model included all features individually, the other applied a feature aggregation strategy. Although the aggregated model was initially expected to perform better, it delivered slightly worse results. However, it provided clearer interpretability regarding the influence of each feature on electricity prices.

The second model differed from the first by only about 1% in terms of error rate, which can be considered an acceptable margin. Given its simplicity, the second model was initially assumed to be more efficient. However, this outcome raises some important considerations. First, the fact that both models underperformed compared to the first model presents a challenge. Future improvements should focus on applying feature aggregation or feature elimination techniques to achieve both a simpler and more accurate model.

Additionally, some of the features used in the models exhibit high correlation with each other, which may lead to redundancy. Reducing multicollinearity and introducing alternative variables may help better capture the influence of distinct factors on electricity prices. Moreover, while this study relied on multiple linear regression, testing other simple mathematical models could provide further insights into prediction performance and model robustness.