Business project: energy price forecasting

Esteban Ortega Dominguez...

Contents

1. Introduction

2. Dataset

2.1 Loading data

```
# Specify the path to your Excel file
excel_file_path <- "data/generation_monthly.xlsx"

column_names <- c("year", "month", "state", "type_of_producer", "energy_source", "generation_mwh")

generation_data <- data.frame()

# Get the list of sheet names in the Excel file
sheet_names <- excel_sheets(excel_file_path)

# Loop through each sheet
for (sheet_name in sheet_names) {
  # Read data from the current sheet, starting from the third row
  data <- read_excel(excel_file_path, sheet = sheet_name, skip = 1)
  colnames(data) <- column_names

  # Combine the data into the 'generation_data' dataframe
  generation_data <- bind_rows(generation_data, data)}

head(generation_data)
```

```
##   year month state                      type_of_producer
## 1 2001     1    AK          Total Electric Power Industry
## 2 2001     1    AK          Total Electric Power Industry
## 3 2001     1    AK          Total Electric Power Industry
```

```
## 4 2001     1    AK           Total Electric Power Industry
## 5 2001     1    AK           Total Electric Power Industry
## 6 2001     1    AK Electric Generators, Electric Utilities
##                 energy_source generation_mwh
## 1                   Petroleum          71085
## 2                 Natural Gas         367521
## 3 Hydroelectric Conventional          104549
## 4                        Wind              87
## 5                       Total         590145
## 6                        Coal          18410
```

```
#Convert chars to factors
convert_chars_to_factors <- function(df) {
  for (col in names(df)) {
    if (is.character(df[[col]])) {
      df[[col]] <- factor(df[[col]])
    }
  }
  return(df)
}

generation_data <- convert_chars_to_factors(generation_data)
generation_data$month <- factor(generation_data$month)
generation_data$year <- factor(generation_data$year)

# Create date column
generation_data$date <- as.Date(paste(generation_data$year, generation_data$month, "01", sep = "-"))
#generation_data$date <- factor(generation_data$date)
```

```
summary(generation_data)
```

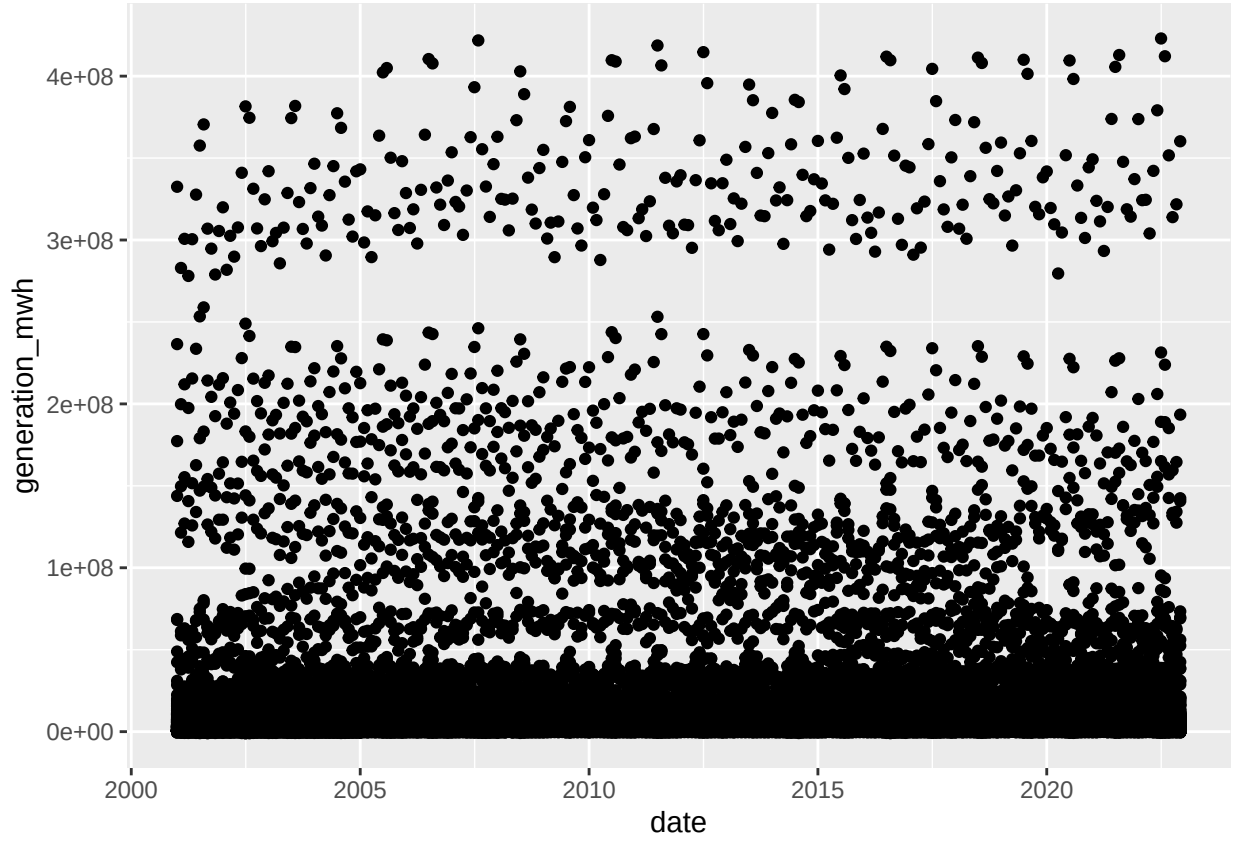
```
##       year            month            state
##  2022   : 25493   12     : 42785   CA     : 15928
##  2020   : 25449   11     : 42728   MI     : 14595
##  2019   : 25426   10     : 42713   PA     : 13593
##  2021   : 25401   9      : 42697   NY     : 13279
##  2018   : 25212   8      : 42664   MN     : 12766
##  2017   : 24999   7      : 42641   NC     : 12718
##  (Other):359666   (Other):255418   (Other):428767
##                                          type_of_producer
##  Combined Heat and Power, Commercial Power       : 55781
##  Combined Heat and Power, Electric Power         : 48498
##  Combined Heat and Power, Industrial Power       : 80070
##  Electric Generators, Electric Utilities         : 96555
##  Electric Generators, Independent Power Producers: 92306
##  Total Electric Power Industry                   :138436
##
##                    energy_source    generation_mwh           date
##  Total                    : 75392   Min.    :  -997855   Min.   :2001-01-01
##  Natural Gas              : 68479   1st Qu.:     1610   1st Qu.:2007-01-01
##  Petroleum                : 64866   Median :    23134   Median :2012-09-01
##  Coal                     : 48845   Mean   :  1392629   Mean   :2012-05-29
##  Other Biomass            : 47531   3rd Qu.:   278528   3rd Qu.:2017-12-01
```

```
##  Hydroelectric Conventional: 39937   Max.   :422975653   Max.   :2022-12-01
##  (Other)                   :166596
```

```
str(generation_data)
```

```
## 'data.frame':    511646 obs. of  7 variables:
##  $ year            : Factor w/ 22 levels "2001","2002",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ month           : Factor w/ 12 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ state           : Factor w/ 53 levels "AK","AL","AR",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ type_of_producer: Factor w/ 6 levels "Combined Heat and Power, Commercial Power",..: 6 6 6 6 6 4 4
##  $ energy_source   : Factor w/ 14 levels "Coal","Geothermal",..: 9 4 3 13 12 1 9 4 3 13 ...
##  $ generation_mwh  : num  71085 367521 104549 87 590145 ...
##  $ date            : Date, format: "2001-01-01" "2001-01-01" ...
```

```
ggplot(data = generation_data, mapping = aes(x = date, y = generation_mwh))+
  geom_point()
```




```
plot_data <- generation_data %>%
  group_by(date, energy_source) %>%
  summarise(gen_by_energy = sum(generation_mwh))
```

```
## `summarise()` has grouped output by 'date'. You can override using the
## `.groups` argument.
```

```
plot_data
```

```
## # A tibble: 3,696 x 3
## # Groups:   date [264]
##    date       energy_source                gen_by_energy
##    <date>     <fct>                                <dbl>
##  1 2001-01-01 Coal                         709101541
##  2 2001-01-01 Geothermal                     4917300
##  3 2001-01-01 Hydroelectric Conventional    75408194
##  4 2001-01-01 Natural Gas                  169554652
##  5 2001-01-01 Nuclear                      274828308
##  6 2001-01-01 Other                          3966680
##  7 2001-01-01 Other Biomass                  4834584
##  8 2001-01-01 Other Gases                    2873768
##  9 2001-01-01 Petroleum                     72447372
## 10 2001-01-01 Pumped Storage                -2354504
## # i 3,686 more rows
```

```
ggplot(data = plot_data, aes(x = date, y = gen_by_energy, color= energy_source)) +
  geom_point()
```