Power Plants

Jeremiah W

November 2020

```
# regarding 'readthedown' theme
# https://cran.r-project.org/web/packages/rmdformats/vignettes/introduction.html

# https://datatables.net/reference/option/
options(DT.options = list(scrollX = TRUE, pagin = TRUE, fixedHeader = TRUE, searchHighlight = TRUE))
```

Introduction

Check out this Kaggle

This data has been gathered at two solar power plants in India over a 34 day period. It has two pairs of files - each pair has one power generation dataset and one sensor readings dataset. The power generation datasets are gathered at the inverter level - each inverter has multiple lines of solar panels attached to it. The sensor data is gathered at a plant level - single array of sensors optimally placed at the plant.

There are a few areas of concern at the solar power plant -

1. Can we predict the power generation for next couple of days? - this allows for better grid management
2. Can we identify the need for panel cleaning/maintenance?
3. Can we identify faulty or suboptimally performing equipment?

Get and Split Data: Generation: Plant 1

```
p1.gd = read_csv('Plant_1_Generation_Data.csv') %>% head(200)
```

```
##
## -- Column specification --------------------------------------
## cols(
##   DATE_TIME = col_character(),
##   PLANT_ID = col_double(),
##   SOURCE_KEY = col_character(),
##   DC_POWER = col_double(),
##   AC_POWER = col_double(),
##   DAILY_YIELD = col_double(),
##   TOTAL_YIELD = col_double()
## )
```

```
p1.gd = read_csv('Plant_1_Generation_Data.csv') %>%
  clean_names() %>%  #lowercase
```

```
  select(sort(tidyselect::peek_vars())) %>% #sort cols alphabetically
  select(where(is.factor),where(is.character),where(is.numeric)) #group by data type
```

```
##
## -- Column specification --------------------------------------
## cols(
##   DATE_TIME = col_character(),
##   PLANT_ID = col_double(),
##   SOURCE_KEY = col_character(),
##   DC_POWER = col_double(),
##   AC_POWER = col_double(),
##   DAILY_YIELD = col_double(),
##   TOTAL_YIELD = col_double()
## )
```

```
#OlsonNames()
#https://stackoverflow.com/questions/41479008/what-is-the-correct-tz-database-time-zone-for-india
```

```
p1.gd$date_time = as.POSIXct(strptime(p1.gd$date_time, "%d-%m-%Y %H:%M"), tz = 'Asia/Kolkata')
```

```
p1.gd %>% head
```

```
## # A tibble: 6 x 7
##   date_time           source_key ac_power daily_yield dc_power plant_id
##   <dttm>              <chr>         <dbl>       <dbl>    <dbl>    <dbl>
## 1 2020-05-15 00:00:00 1BY6WEcLG~        0           0        0  4135001
## 2 2020-05-15 00:00:00 1IF53ai7X~        0           0        0  4135001
## 3 2020-05-15 00:00:00 3PZuoBAID~        0           0        0  4135001
## 4 2020-05-15 00:00:00 7JYdWkrLS~        0           0        0  4135001
## 5 2020-05-15 00:00:00 McdE0feGg~        0           0        0  4135001
## 6 2020-05-15 00:00:00 VHMLBKoKg~        0           0        0  4135001
## # ... with 1 more variable: total_yield <dbl>
```

```
p1.gd %>% slice_sample(n = 5)
```

```
## # A tibble: 5 x 7
##   date_time           source_key ac_power daily_yield dc_power plant_id
##   <dttm>              <chr>         <dbl>       <dbl>    <dbl>    <dbl>
## 1 2020-05-15 06:15:00 rGa61gmuv~    30.5        3        315.   4135001
## 2 2020-06-14 06:00:00 WRmjgnKYA~     5.24       0.125     54.5  4135001
## 3 2020-06-11 22:00:00 z9Y9gH1T5~     0       5599          0    4135001
## 4 2020-05-28 03:30:00 z9Y9gH1T5~     0          0          0    4135001
## 5 2020-06-13 10:45:00 YxYtjZvoo~   896.      2607.      9176.   4135001
## # ... with 1 more variable: total_yield <dbl>
```

```
p1.gd %>% glimpse()
```

```
## Rows: 68,778
## Columns: 7
## $ date_time   <dttm> 2020-05-15 00:00:00, 2020-05-15 00:00:00, 2020-05-15 0...
## $ source_key  <chr> "1BY6WEcLGh8j5v7", "1IF53ai7Xc0U56Y", "3PZuoBAID5Wc2HD"...
## $ ac_power    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ daily_yield <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ dc_power    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ plant_id    <dbl> 4135001, 4135001, 4135001, 4135001, 4135001, 4135001, 4...
## $ total_yield <dbl> 6259559, 6183645, 6987759, 7602960, 7158964, 7206408, 7...
```

EDA