EDA

Margaret Reed

2023-05-07

```
library(tidyverse)
library(lubridate)
library(prophet)
library(MLmetrics)
library(scales)
```

```
data <- read_csv("../raw_data/weekly_usage.csv")
```

Exploratory data analysis

data cleaning

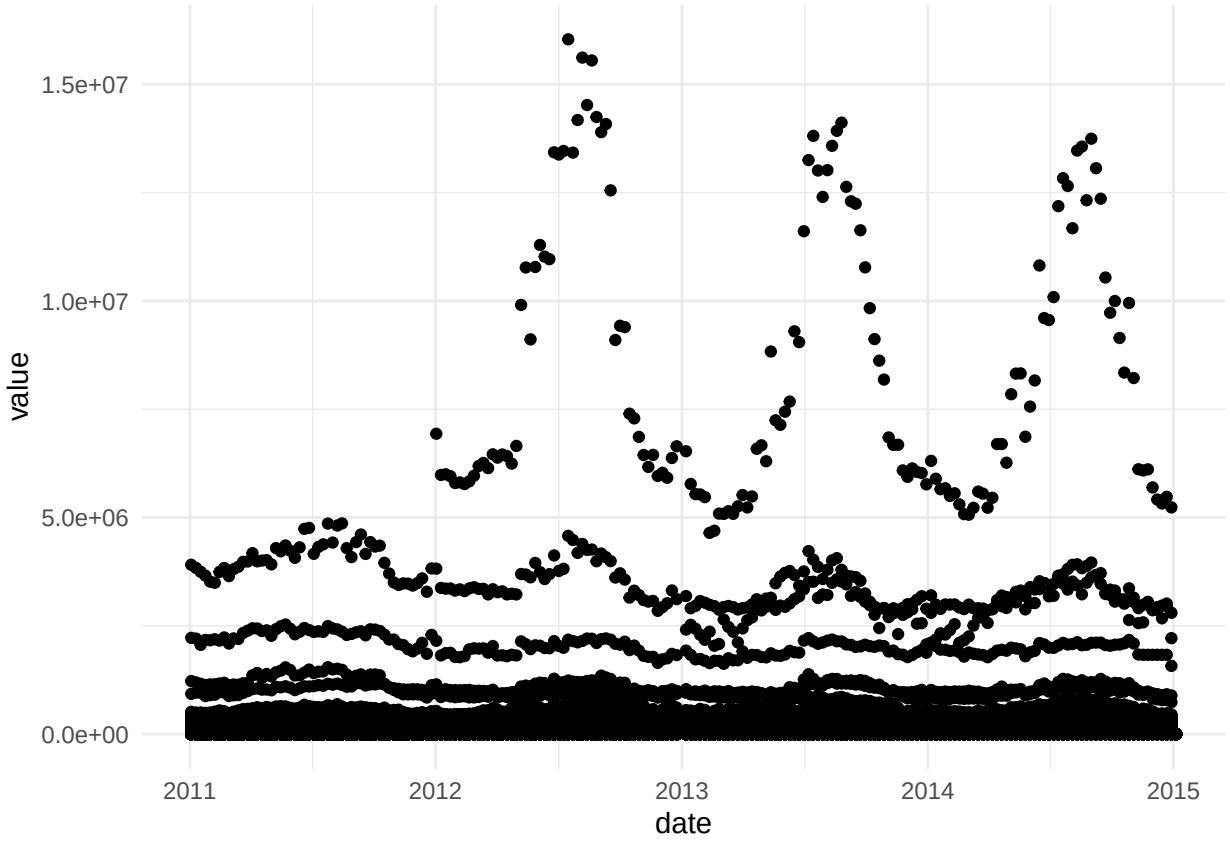
To clean the data, all that was really required was creating a date object with the year and week variables.

```
data <- data %>%
  mutate(
    date = lubridate::parse_date_time(paste(year, week, 1, sep="/"),'Y/W/w')
  ) %>%
  select(date, account, value, t)
```

plotting the data

First I plotted the data as it was.

```
data %>%
   ggplot(aes(x = date, y = value)) +
   geom_point() +
  theme_minimal()
```



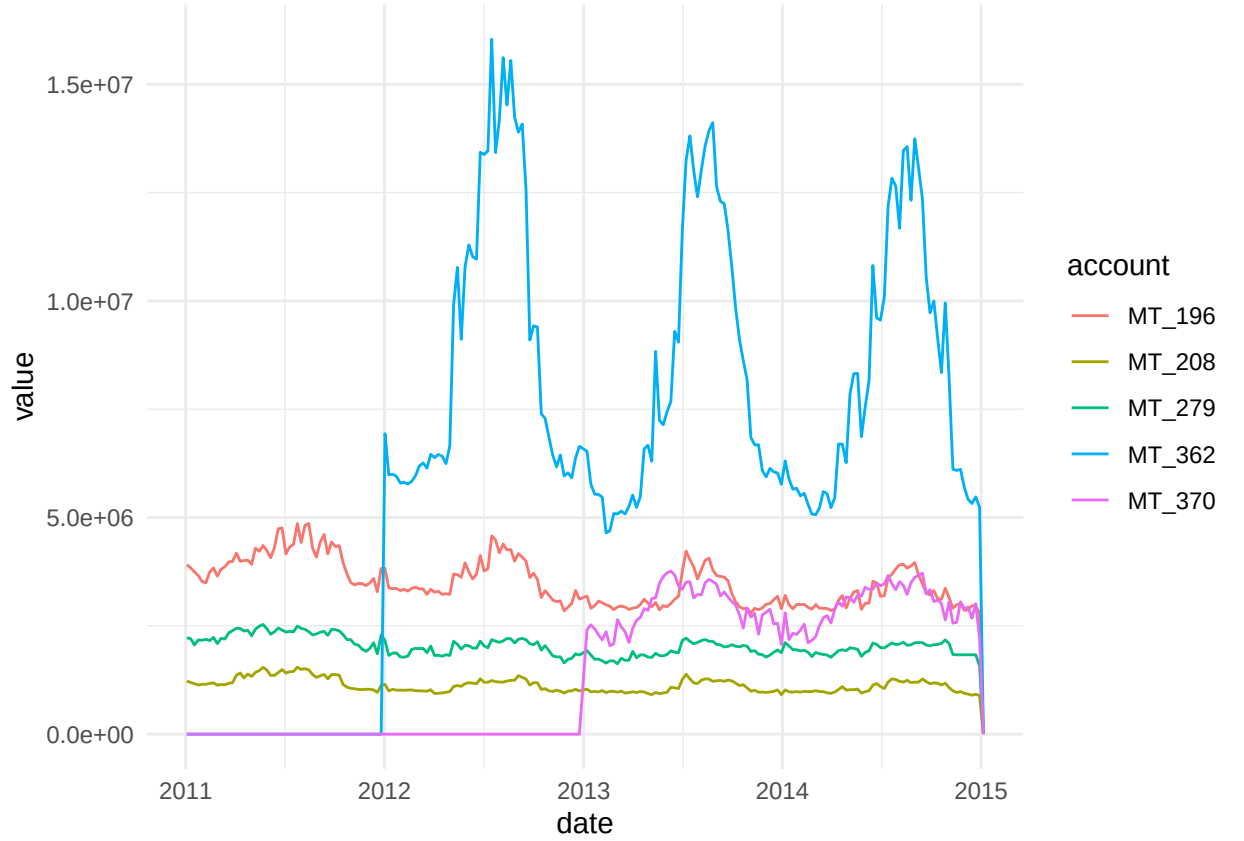
It looks pretty messy but there are clearly several different trends within the data.

finding trends

To get a better look at these trends I decided to look at the 5 top consumers of electricity among the different accounts

```
top_5_accounts <- data %>%
  group_by(account) %>%
  summarize(
    total_val = sum(value)
  ) %>%
  arrange(desc(total_val)) %>%
  slice(1:5) %>%
  pull(account)

data %>%
  filter(
    account %in% top_5_accounts
  ) %>%
   ggplot(aes(x = date, y = value, color = account)) +
   geom_line() +
  theme_minimal()
```

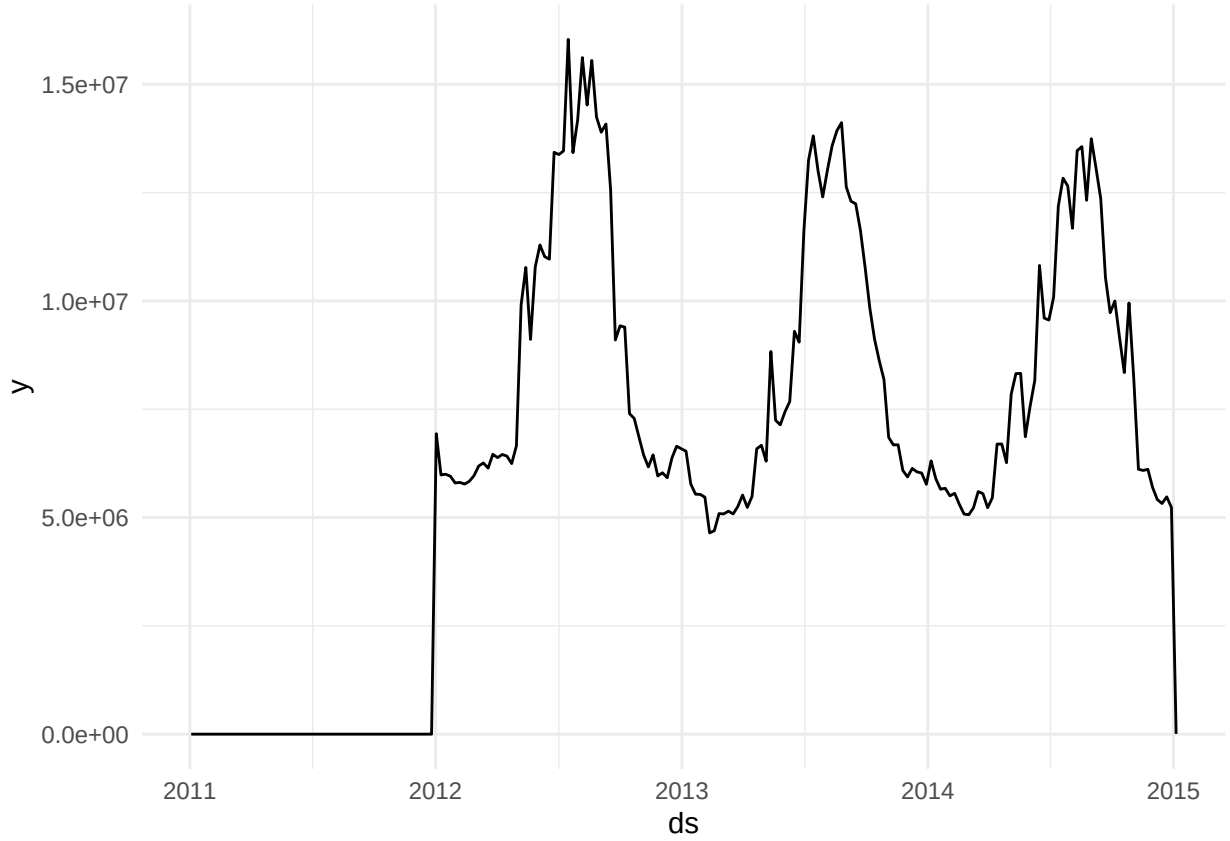


Clearly accounts MT_362, _196, _279, _370, and _208 are the top users and have fairly defined trends with seasonal components.

First I will look at account MT_362

```
MT_362 <- data %>%
  filter(account == "MT_362") %>%
  select(date, value) %>%
  rename(ds = date, y = value)

MT_362 %>%
  ggplot(aes(x = ds, y = y)) +
  geom_line() +
  theme_minimal()
```



There appears to be an outlier at the end of the data so I will remove it for the sake of analysis

```
MT_362 <- MT_362 %>%
  slice(1: nrow(MT_362) - 1)
```

```
n <- nrow(MT_362)
N_train <- round(0.7*n, 0)
N_validation <- round(0.2*n, 0)
N_test <- n - N_train - N_validation

train <- MT_362 %>%
  slice(1:N_train)

validation <- MT_362 %>%
  anti_join(train) %>%
  slice(1: N_validation)

test <- MT_362 %>%
  anti_join(train) %>%
  anti_join(validation)

 m <- prophet(rbind(train, validation))
```

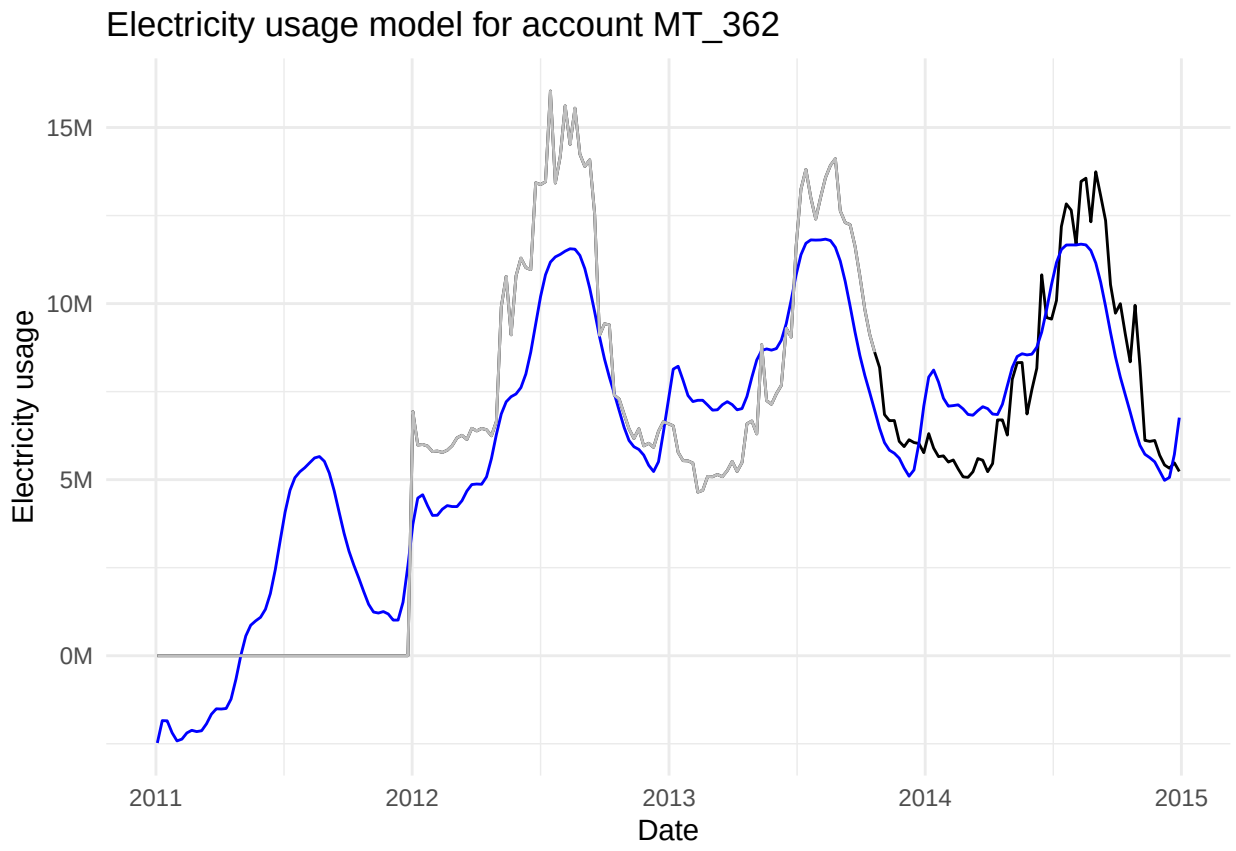
```
future <- make_future_dataframe(m, periods = N_test, freq = "week")
```

```
forecast <- predict(m, future)
```

```
pred_test <- forecast %>%
  slice_tail(n=N_test) %>%
  pull(yhat)

true_test <- test %>%
  pull(y)
```

```
full_join(MT_362, forecast) %>%
  ggplot() +
  geom_line(aes(x = ymd(ds), y = y), color = "black") +
  geom_line(aes(x = ymd(ds), y = yhat), color = "blue") +
  theme_minimal() +
  geom_line(data = MT_362%>%slice_head(n=N_train), aes(x = ymd(ds), y = y), color = "grey") +
  scale_y_continuous(labels = label_number(scale = 0.000001, suffix = "M")) +
  labs(
    x = "Date",
    y = "Electricity usage",
    title = "Electricity usage model for account MT_362"
  )
```




```
s <- round(N_test/3, 0)

MAPE(pred_test, true_test)
```

```
## [1] 0.1508919
```

```
MAPE(pred_test[1:s], true_test[1:s])
```

```
## [1] 0.1479772
```

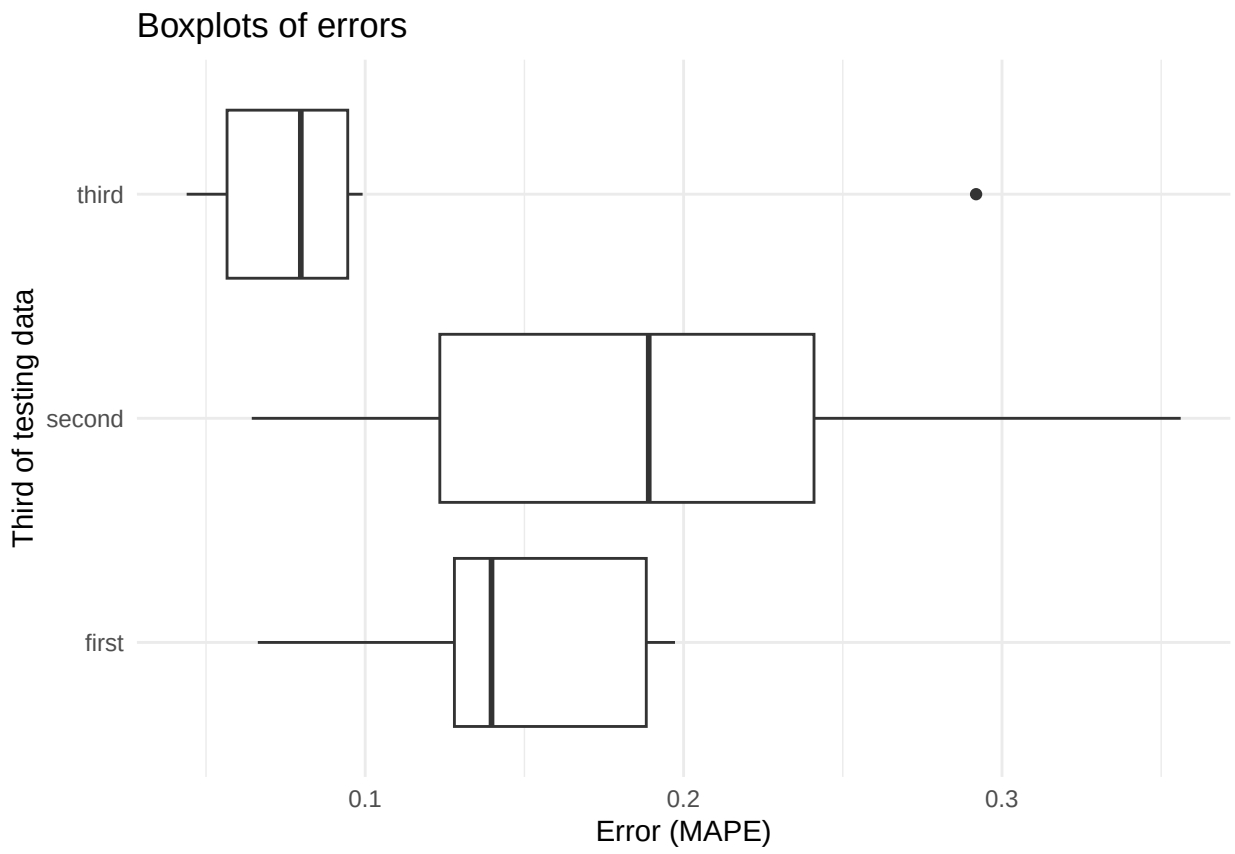
```
MAPE(pred_test[(s+1):(2*s)], true_test[(s+1):(2*s)])
```

```
## [1] 0.1911925
```

```
MAPE(pred_test[(2*s+1):N_test], true_test[(2*s+1):N_test])
```

```
## [1] 0.107275
```

```
tibble(
 first = map2_dbl(pred_test[1:s], true_test[1:s], MAPE),
 second = map2_dbl(pred_test[(s+1):(2*s)], true_test[(s+1):(2*s)], MAPE),
third = map2_dbl(pred_test[(2*s+1):(3*s)], true_test[(2*s+1):(3*s)], MAPE)
) %>%
  pivot_longer(
    cols = everything(),
    names_to = "segment",
    values_to = "error"
  ) %>%
  ggplot(aes(x = error, y = segment)) +
  geom_boxplot() +
  labs(
    x = "Error (MAPE)",
    y = "Third of testing data",
    title = "Boxplots of errors"
  ) +
  theme_minimal()
```


Boxplots of errors
Third of testing data
third
second
first
0.1
0.2
0.3
Error (MAPE)