Report About Individual Household Electric Power Consumption

QIANG SUN/21804416

Goals

1. Explore some obvious rules and patterns from the data.
2. Build a model to predict power consumption for each month.
3. Build a model to figure out the influence on power factor from different sub meters.
4. Find rules in the power consumption.

Manage Data

1. Data frame pre-process

a. Load it as data frame with R language
b. Assign each column with proper type
c. Assign a column to control the version and data provenance

2. Deal with missing data

First, we use summary to check the portion of the missing data for each column.

```
> summary(mydata)
      Date                Time          Global_active_power Global_reactive_power    Voltage      Global_intensity Sub_metering_1
 Min.   :2006-12-16   17:24:00:    1442   Min.   : 0.076    Min.   :0.000         Min.   :223.2   Min.   : 0.200   Min.   : 0.000
 1st Qu.:2007-12-12   17:25:00:    1442   1st Qu.: 0.308    1st Qu.:0.048         1st Qu.:239.0   1st Qu.: 1.400   1st Qu.: 0.000
 Median :2008-12-06   17:26:00:    1442   Median : 0.602    Median :0.100         Median :241.0   Median : 2.600   Median : 0.000
 Mean   :2008-12-05   17:27:00:    1442   Mean   : 1.092    Mean   :0.124         Mean   :240.8   Mean   : 4.628   Mean   : 1.122
 3rd Qu.:2009-12-01   17:28:00:    1442   3rd Qu.: 1.528    3rd Qu.:0.194         3rd Qu.:242.9   3rd Qu.: 6.400   3rd Qu.: 0.000
 Max.   :2010-11-26   17:29:00:    1442   Max.   :11.122    Max.   :1.390         Max.   :254.2   Max.   :48.400   Max.   :88.000
                      (Other) :2066607   NA's   :25979     NA's   :25979         NA's   :25979   NA's   :25979    NA's   :25979
 Sub_metering_2   Sub_metering_3     version      DateFull                      DateTimes
 Min.   : 0.000   Min.   : 0.000   Min.   :1   Min.   :2006-12-16 17:24:00   Min.   :2017-10-01 00:00:00
 1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.:1   1st Qu.:2007-12-12 00:48:30   1st Qu.:2017-10-01 06:00:00
 Median : 0.000   Median : 1.000   Median :1   Median :2008-12-06 08:13:00   Median :2017-10-01 12:00:00
 Mean   : 1.299   Mean   : 6.458   Mean   :1   Mean   :2008-12-06 08:14:40   Mean   :2017-10-01 11:59:32
 3rd Qu.: 1.000   3rd Qu.:17.000   3rd Qu.:1   3rd Qu.:2009-12-01 14:37:30   3rd Qu.:2017-10-01 18:00:00
 Max.   :80.000   Max.   :31.000   Max.   :1   Max.   :2010-11-26 21:02:00   Max.   :2017-10-01 23:59:00
 NA's   :25979    NA's   :25979                NA's   :120

[1] "Missing data rate for each column"
                 Date                  Time   Global_active_power Global_reactive_power               Voltage      Global_intensity
          0.000000000           0.000000000           1.251843746           1.251843746           1.251843746           1.251843746
       Sub_metering_1        Sub_metering_2        Sub_metering_3               version              DateFull             DateTimes
          1.251843746           1.251843746           1.251843746           0.000000000           0.005782411           0.000000000
        Date Time version DateTimes DateFull Global_active_power Global_reactive_power Voltage Global_intensity Sub_metering_1
2049160    1    1       1         1        1                   1                     1       1                1              1
120        1    1       1         1        0                   1                     1       1                1              1
25979      1    1       1         1        1                   0                     0       0                0              0
           0    0       0         0      120               25979                 25979   25979            25979          25979
        Sub_metering_2 Sub_metering_3
2049160              1              1      0
120                  1              1      1
25979                0              0      7
                 25979          25979 181973

                 Date                  Time   Global_active_power Global_reactive_power               Voltage      Global_intensity
          0.000000000           0.000000000           1.251843746           1.251843746           1.251843746           1.251843746
       Sub_metering_1        Sub_metering_2        Sub_metering_3               version              DateFull             DateTimes
          1.251843746           1.251843746           1.251843746           0.000000000           0.005782411           0.000000000
```

Fig 1: Summary of raw data

As seen in the above figures. We can see there 120 rows of missing data with Date, and 25979 rows missing with other 7 columns. Properly around 1.25% of the whole data set.

Then we examine the whole data set to check the distribution of missing data and the whole data set. We need to find whether the missing data is random or not, and find a solution for the missing data. We can see it from the Fig 2

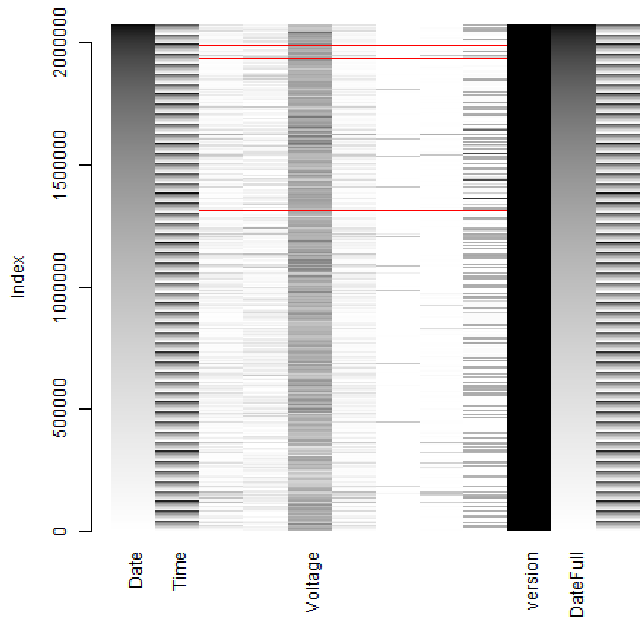


Fig 2.1 The missing data and the distribution of all the data.

The missing data is in red, the colour depth represents the value of the data.

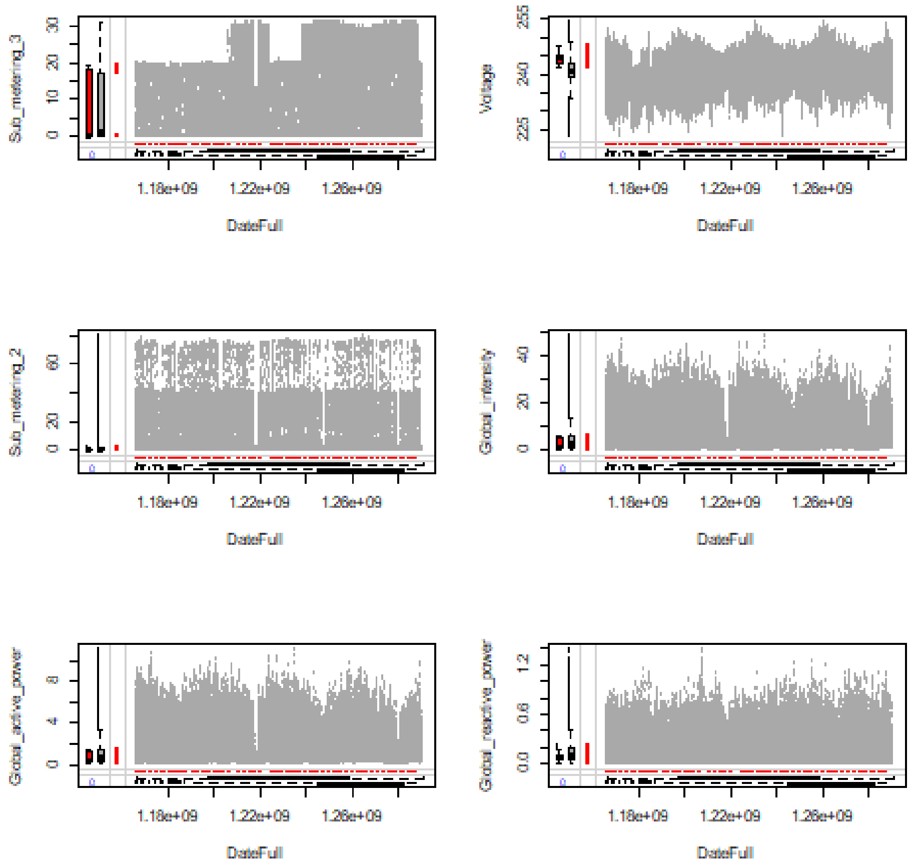


Fig 2.2 The distribution of Missing data

We can find that the missing data of time are mostly laying back of the whole date, which most in 2007 and 2008. And the distribution of the other data is mostly the same distribution as the original ones.

After analysing all the missing data, we can infer that the missing data may be caused by massive outage. And it will not affect the whole distribution of the data set if we drop them all. And we have no way but dropping all the missing data.

Explore Data

1. Data Process

a. Add the column: power consumption which not recorded by the 3 meters.
b. Separate the date for Year, Month, Day, Hour, Minutes, and Day of Year
c. Calculate apparent power, and record it in Global_Power
d. Calculate power factor, and record it in Phase

2. Explore the data range

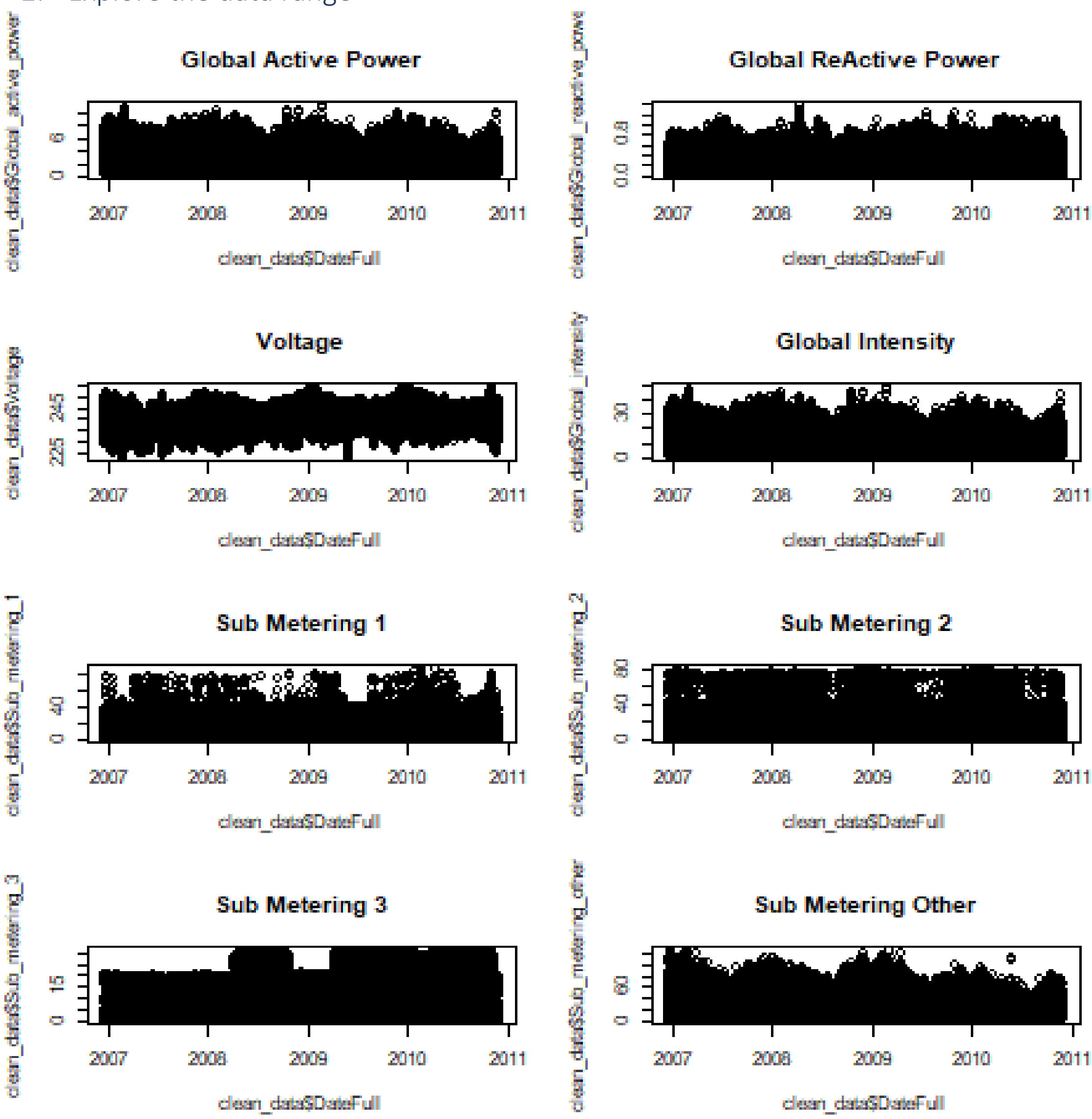


Fig 3.1 Data Range for all the data columns

We can infer from the Fig 3.1 that, the power consumption should be the lowest in June, and highest in December. Maybe the household was out for holiday in every June. And we can infer that the household is living in Southern Hemisphere. Voltage is distributed between 210-260. And there are some obvious patterns from the 3 meters. The summary of cleaned data supports our inferences.

```
> summary(clean_data)
      Date                 Time          Global_active_power Global_reactive_power    Voltage      Global_intensity Sub_metering_1   Sub_metering_2
 Min.   :2006-12-16   19:51:00:    1427   Min.   : 0.076      Min.   :0.0000        Min.   :223.2   Min.   : 0.200   Min.   : 0.000   Min.   : 0.000
 1st Qu.:2007-12-10   19:52:00:    1427   1st Qu.: 0.308      1st Qu.:0.0480        1st Qu.:239.0   1st Qu.: 1.400   1st Qu.: 0.000   1st Qu.: 0.000
 Median :2008-11-30   19:53:00:    1427   Median : 0.602      Median :0.1000        Median :241.0   Median : 2.600   Median : 0.000   Median : 0.000
 Mean   :2008-12-01   19:54:00:    1427   Mean   : 1.092      Mean   :0.1237        Mean   :240.8   Mean   : 4.628   Mean   : 1.122   Mean   : 1.299
 3rd Qu.:2009-11-23   19:55:00:    1427   3rd Qu.: 1.528      3rd Qu.:0.1940        3rd Qu.:242.9   3rd Qu.: 6.400   3rd Qu.: 0.000   3rd Qu.: 1.000
 Max.   :2010-11-26   19:56:00:    1427   Max.   :11.122      Max.   :1.3900        Max.   :254.2   Max.   :48.400   Max.   :88.000   Max.   :80.000
                      (Other) :2040598
 Sub_metering_3      version       DateFull                        DateTimes                     Sub_metering_other      Year               Month
 Min.   : 0.000   Min.   :3   Min.   :2006-12-16 17:24:00   Min.   :2017-10-01 00:00:00   Min.   :  0.000     Length:2049160     Length:2049160
 1st Qu.: 0.000   1st Qu.:3   1st Qu.:2007-12-10 06:07:45   1st Qu.:2017-10-01 06:00:00   1st Qu.:  3.800     Class :character   Class :character
 Median : 1.000   Median :3   Median :2008-11-30 02:22:30   Median :2017-10-01 12:00:00   Median :  5.500     Mode  :character   Mode  :character
 Mean   : 6.459   Mean   :3   Mean   :2008-12-02 02:01:04   Mean   :2017-10-01 11:59:46   Mean   :  9.315
 3rd Qu.:17.000   3rd Qu.:3   3rd Qu.:2009-11-23 21:01:15   3rd Qu.:2017-10-01 18:00:00   3rd Qu.: 10.367
 Max.   :31.000   Max.   :3   Max.   :2010-11-26 21:02:00   Max.   :2017-10-01 23:59:00   Max.   :124.833

     Day                 Hour                Minute              DayOfYear            Global_Power          Phase
 Length:2049160      Length:2049160      Length:2049160      Length:2049160      Min.   : 0.0760   Min.   :0.5559
 Class :character    Class :character    Class :character    Class :character    1st Qu.: 0.3319   1st Qu.:0.9520
 Mode  :character    Mode  :character    Mode  :character    Mode  :character    Median : 0.6340   Median :0.9934
                                                                                 Mean   : 1.1096   Mean   :0.9637
                                                                                 3rd Qu.: 1.5385   3rd Qu.:0.9997
                                                                                 Max.   :11.1234   Max.   :1.0000
```

Fig 3.2 Summary of Data which have been cleaned.

3. Some Interesting Patterns

a. Voltage and Intensity VS Power Factor

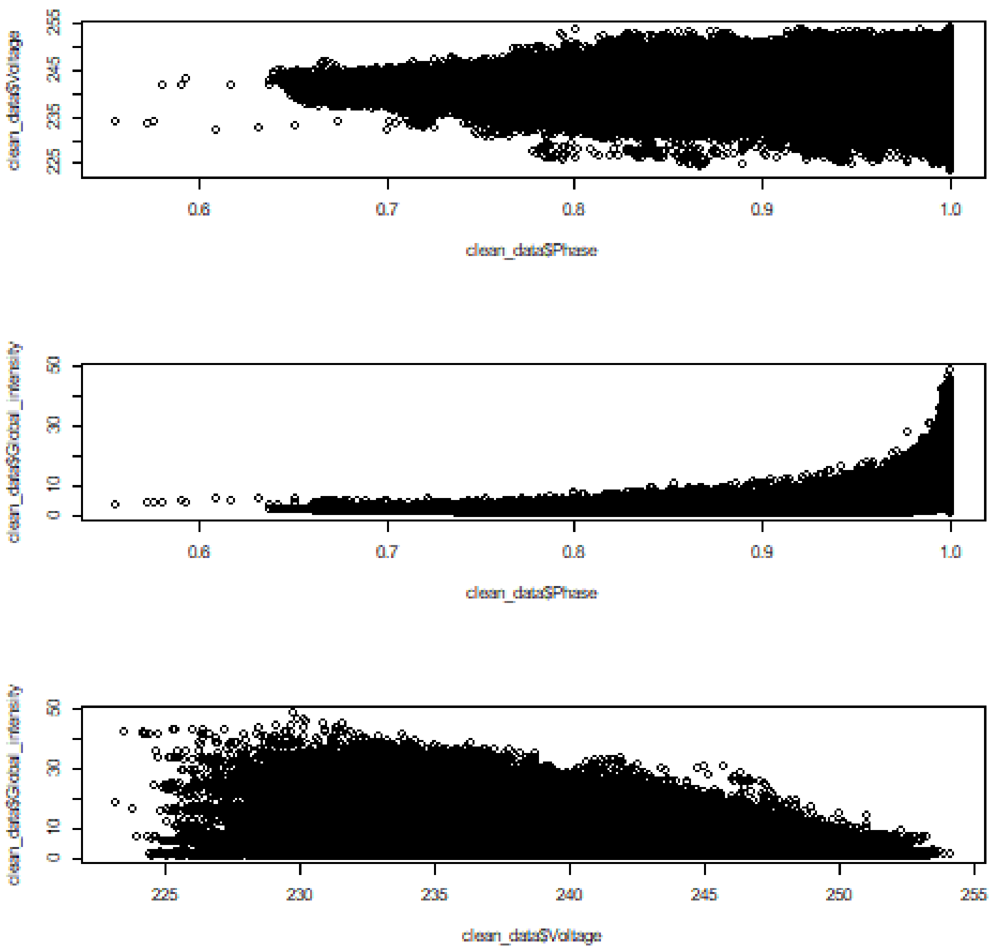


Fig 4.1 Some patterns between voltage, intensity and power factor

The distribution for Power Factor and Voltage is normal distribution, and the distribution for Power Factor and Intensity is exponential distribution. These 2 patterns can be useful for power providers to provide stable voltage and better electric quality, or used to prevent intensity to be too high, which could cause fire or grid crash.

b. Proportion of different meters

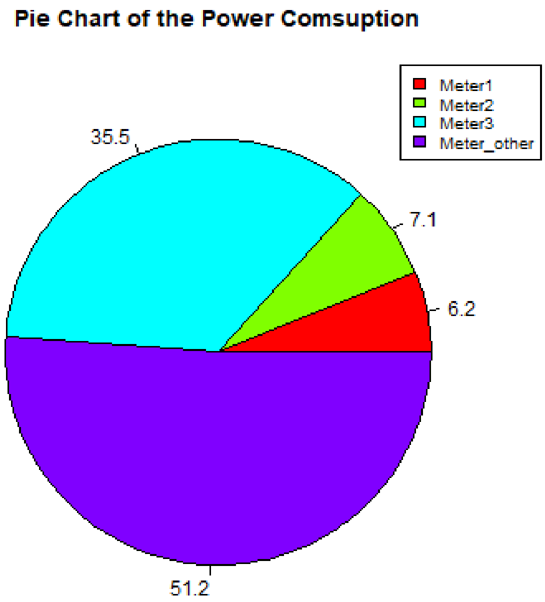


Fig 4.2 Proportion of different metters

Most power are consumed by the other equipment, and around half of the power are consumed by the 3 meters, Air Condition and Heater consume most of the power.

If the household want to save their cost of electric, they should decrease the usage of AC and heater, or they should replace a new equipment which consume less power.

Build Models

1. Regression: No Linear

We want to predict the power consumption for days or months, so we can predict the cost of electric for the household in the future.

To fulfil that, we first group the power consumption data with day and month. And do the boxplots to check which version is better to fitting.

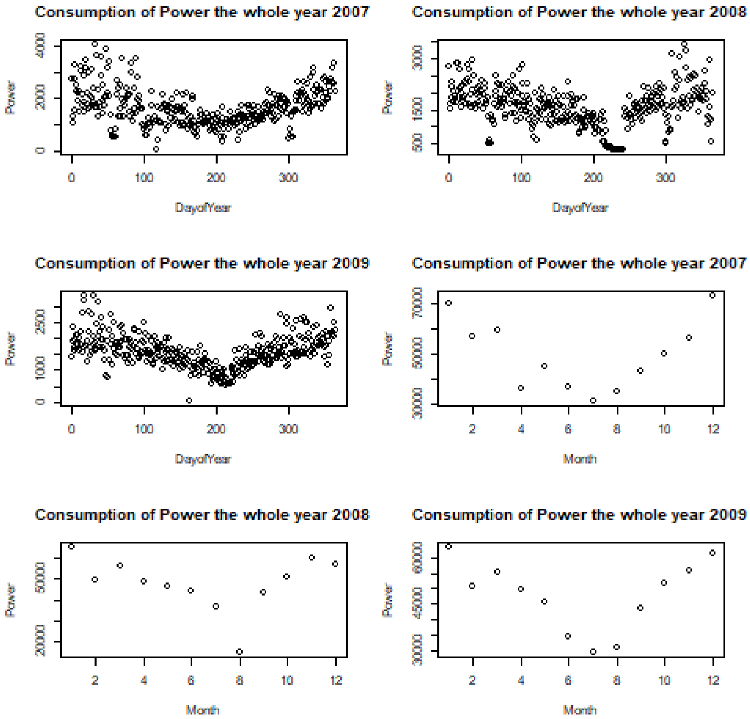


Fig 5.1 Scatter Plots of Month and Day in 2007,2008,2009

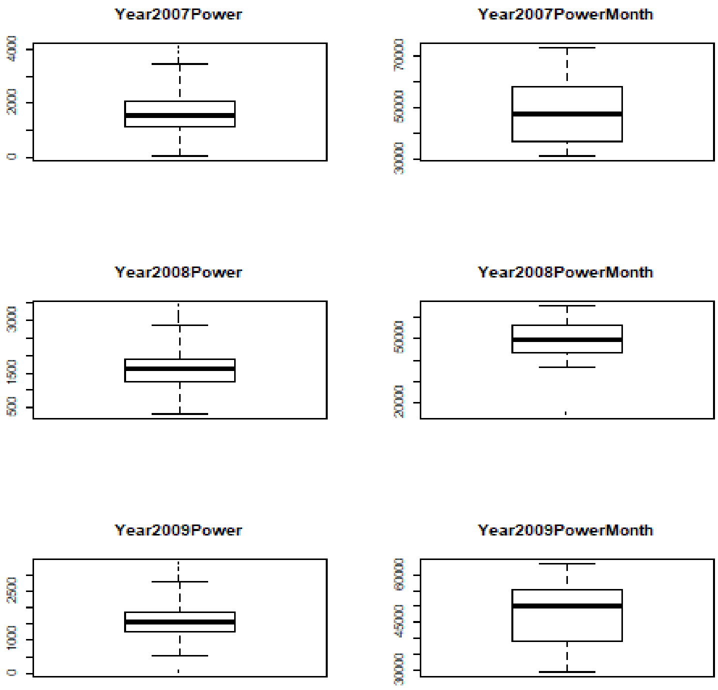


Fig 5.2 Boxplots for Power Consumption

After examining the boxplots and scatter plots, we decide to fit month vs Power Consumption with no linear regression.

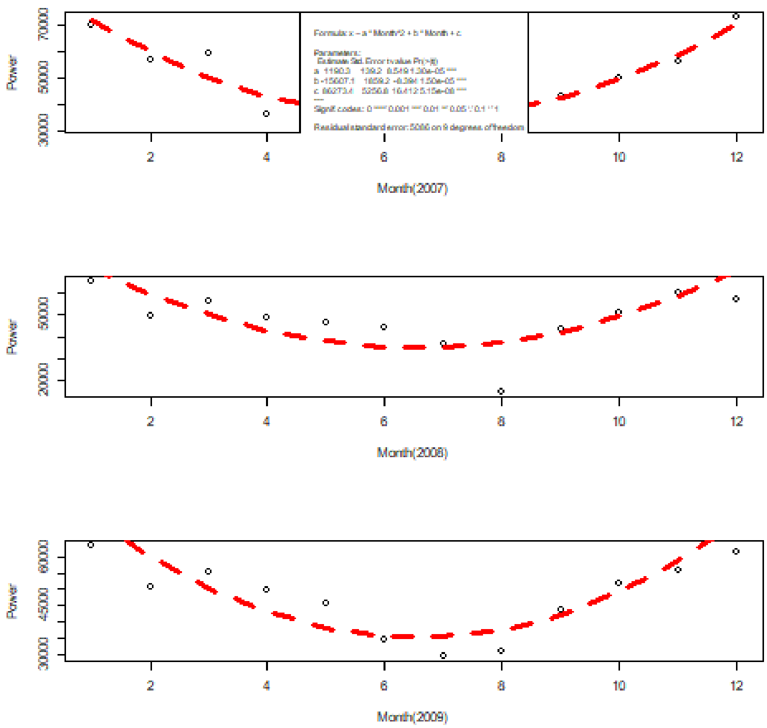


Fig 5.3 Fitting Plot

It fits pretty good. The equation is: $\mathrm{Power} = 1190.3 \times Month^2 - 156007.1 * Month + 86273.4$

2. Regression: Multivariable Linear

The higher the factor power is, the less power waste, so we want to find out which sub meter influence the power factor most. With that power provider can

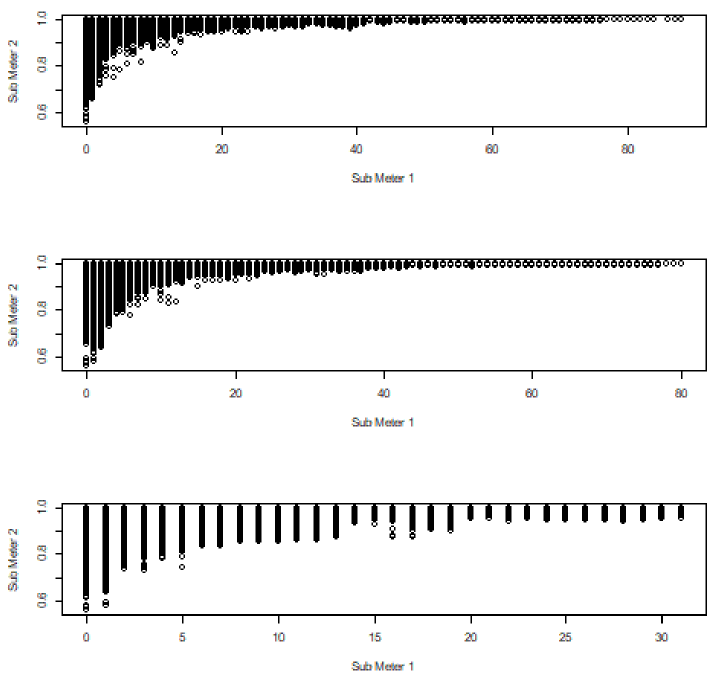
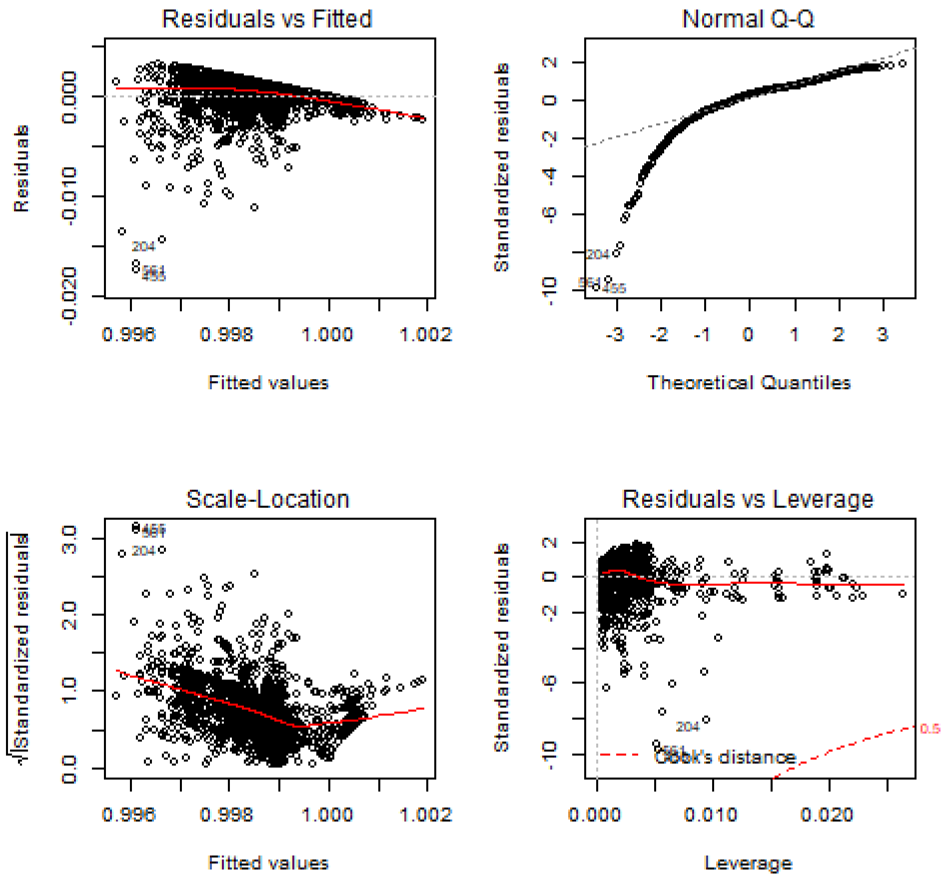


Fig 6.1 Sub Meters VS Power Factor

Except when the sub meter power is smaller than 5, the power sub meters consumed is basically linear with power factor. So, we do the multivariable regression.




```
Call:
lm(formula = build_model_2_bt_sample$clean_data.Phase ~ build_model_2_bt_sample$clean_data.Sub_metering_1 +
    build_model_2_bt_sample$clean_data.Sub_metering_2 + build_model_2_bt_sample$clean_data.Sub_metering_3)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0175184 -0.0006204  0.0004065  0.0010162  0.0033315 

Coefficients:
                                                   Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                                       9.938e-01  4.422e-04 2247.578  < 2e-16 ***
build_model_2_bt_sample$clean_data.Sub_metering_1 6.735e-05  4.124e-06   16.333  < 2e-16 ***
build_model_2_bt_sample$clean_data.Sub_metering_2 4.318e-05  2.831e-06   15.251  < 2e-16 ***
build_model_2_bt_sample$clean_data.Sub_metering_3 6.543e-05  2.246e-05    2.913  0.00362 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.001782 on 1996 degrees of freedom
Multiple R-squared:  0.188,	Adjusted R-squared:  0.1868 
F-statistic:   154 on 3 and 1996 DF,  p-value: < 2.2e-16
```

Fig 6.2 Fitting Multivariable Regression

We can see that the fitting is basically ok. Meter 1 contributes the most, and then the Meter 3. Meter 2 contributes the least for power factor.

3. Association Rules

Meter 1 represents kitchen with Dishwasher, oven and microwave. Meter with washing machine, tumble-drier, and refrigerator, Meter 3 measures the power consumption from Air Conditioner and Heater. We want to find patterns between different usage. For example, whether the household prefer to use the washing machine after they use microwave to cook in afternoon. When we know about this thing, the owner of the household can combine the rules with the electric price. Peak and valley electric charges are applied somewhere. So, owner can find a way to adjust their custom to use the washing machine in mid night, which can lead to reduced cost.

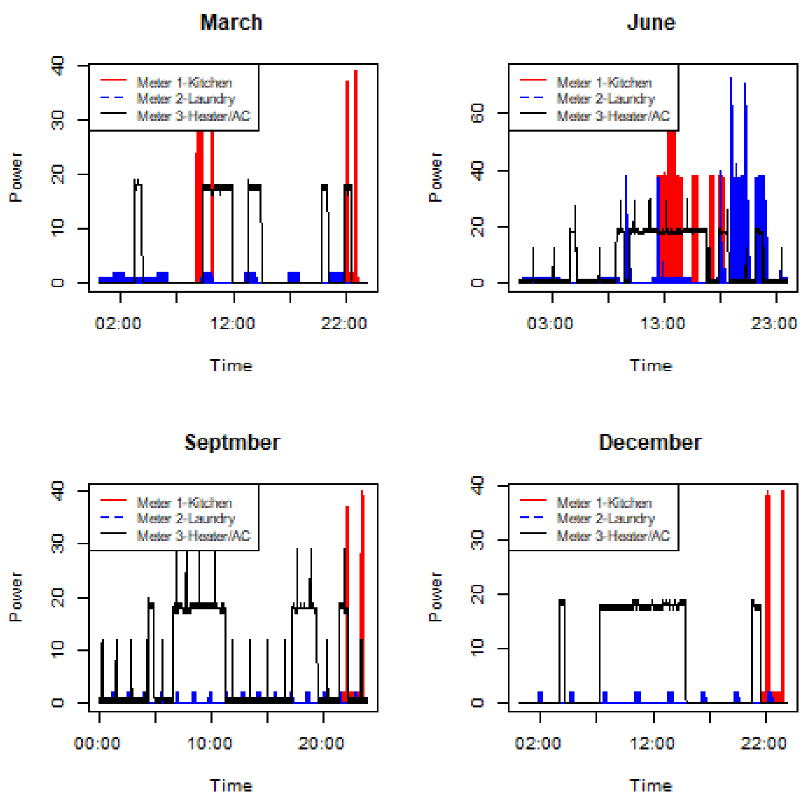


Fig 7.1 Examples of Power Consumption for a Day

We can see that different usage behaviours have different threshold value which can be reflected from the Fig 7.1. After doing some research, we can transform the data into different usage pattern with the conditions below.

Table 1.1: Conditions to transform data into proper format for Association Rule Mining

Mark	Behaviour	Power Range(W*h)	Least Duration(min)
A	Dishwasher/ Microwave	31-50	25
B	Oven	>50	45
C	Refrigerator	0-10	480
D	Washing Machine/ Drier	11-50	35
E	Washing Machine + Drier	>50	35
F	Heater	11-35	80
G	AC	16-25	480
H	Heater + AC	>25	480

Within the least duration, the same behaviour will be treated as the same. The same behaviour happens the first time in the day will be marked as ‘A1’, the second will be marked as ‘A2’. To transform the data into the format we want. First, we find the behaviour which meets the requirements for A-H separately and store with timestamp. Then we sort all the behaviour by time, and extract the behaviour for each day. Then we use Apriori with the conditions we set to find the rules.

```
     lhs              rhs  support   confidence lift
[1]  {A1,A2}      => {F1} 0.6712707 0.9346154  1.0099426
```

Fig 7.2 One outcome from association rules mining

The Fig 7.2 means, the household usually open their Heater after they use the kitchen.

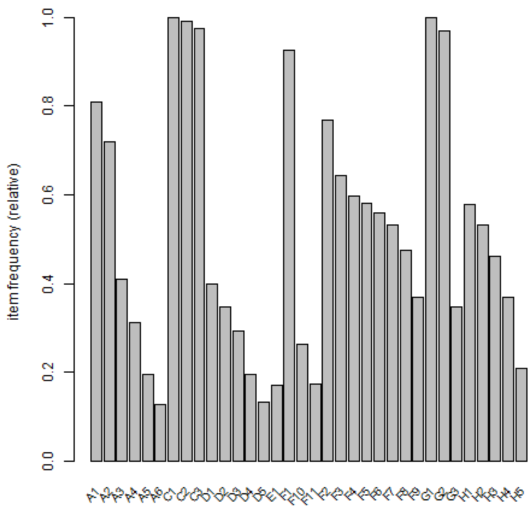


Fig 7.3 Frequency For different behaviour

The Fig 7.3 show the frequency for different behaviour, which we can find that the household didn’t use washing machine frequently, but kept the refrigerator opening.

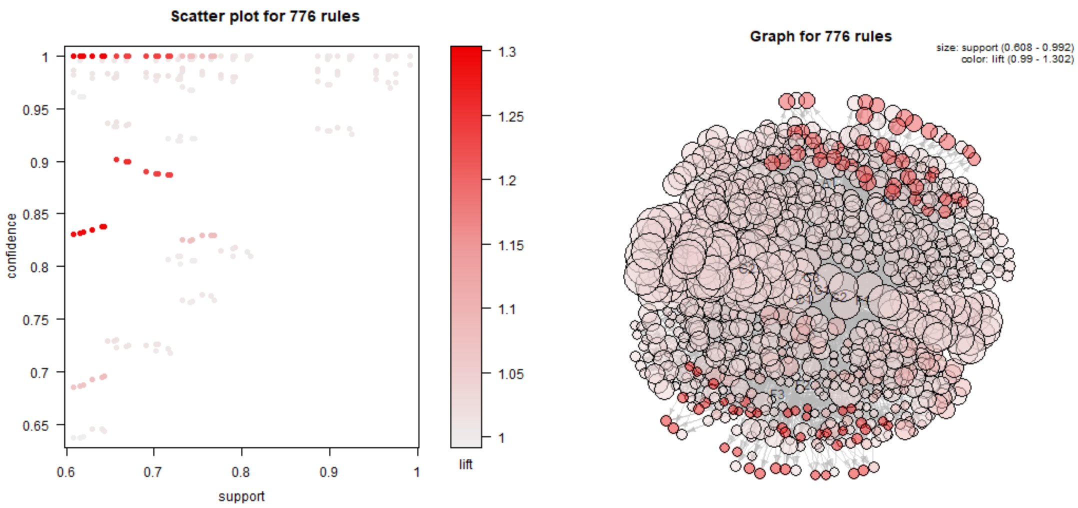


Fig 7.4 Plots for rules

Appendix

Code

Github Repo link: https://github.com/PascalSun/Data_Science_Report