Power Usage Data

C4T1 – Sugitha Devarajan

Agenda

- Background/Objective
- Approach
- Data Management
- Data Description
- Initial Analysis and Findings
- Recommendations
- Q&A

Background/Objective

- Client wants to know if a particular residence was occupied during summer of 2008 by analyzing the power consumption data of the residence.
- **Questions to answer**:
  - What information is contained within the data records?
  - Is the data complete? Is anything missing?
  - What are the typical power usage patterns for this residence?
  - Are these “typical” patterns true for the time period in question?
  - If not, what, if anything, can be used to help support the client’s claims?
  - Are there any outliers or events depicted in the data that may undermine the client’s claims?
  - Are there any recommendations for questions we should be asking?
  - Is any additional information needed from the law firm to conduct the analysis?

Approach

- Understanding Domain - Smart Homes, sub-meters and household power consumption.
- Data munging and sub-setting will be essential to the analytic process.
- Initial data exploration should be used to understand any potential issues, conduct early preprocessing and note summary statistics.
- Deep dive using visualize and analyze energy data to answer the questions in the objective.

Data Management

- Data management is an administrative process that includes acquiring, validating, storing, protecting, and processing required data to ensure the accessibility, reliability, and timeliness of the data.

What is involved in a complete data management model?

- **Data governance**, which is the planning of all aspects of data management. This commonly includes ensuring availability, usability, consistency, integrity, and security of data managed by an organization.
- **Data architecture**, or the overall structure of an organization's data and how it fits into a broader enterprise architecture.
- **Data modeling and design**, which covers data analytics and the design, building, testing, and maintenance of analytics systems.
- **Data storage and operations**, which is concerned with the physical hardware used to store and manage data.
- **Data security**, which encompasses all elements of protecting data and ensuring only authorized users have access.
- **Data integration and interoperability**, which includes everything to do with the transformation of data into a structured form (i.e., in an organized database) and the work necessary to maintain it.
- **Documents and content**, which includes all forms of unstructured data and the work necessary to make it accessible to, and integrated with, structured databases.
- **Reference and master data**, or the process of managing data in such a way that redundancy and other mistakes are reduced by standardizing data values.
- **Data warehousing** and business intelligence, which involves the management and application of data for analytics and business decision making.
- **Metadata**, which involves all elements of creating, collecting, organizing, and managing metadata (data that references other data, like headers, etc.).
- **Data quality,** which involves the practices of monitoring data and data sources to ensure quality information is being delivered, integrity is being maintained, and poor-quality data is being filtered out.

All of these elements have to be included in a total data management model;

Data Description

- Measurements of electric power consumption in one household with a one-minute sampling rate over a period of almost 4 years. Different electrical quantities and some sub-metering values are available.
- Data contains measurements gathered in a house located in Sceaux (7km of Paris, France) between December 2006 and November 2010 (47 months).
- **Attribute Information:**

  1. **Date**: Date in format dd/mm/yyyy
  2. **Time**: time in format hh:mm:ss
  3. **Global_active_power**: household global minute-averaged active power (in kilowatt)
  4. **Global_reactive_power**: household global minute-averaged reactive power (in kilowatt)
  5. **Goltage**: minute-averaged voltage (in volt)
  6. **Global_intensity**: household global minute-averaged current intensity (in ampere)
  7. **Sub_metering_1**: energy sub-metering No. 1 (in watt-hour of active energy). It corresponds to the kitchen, containing mainly a dishwasher, an oven and a microwave (hot plates are not electric but gas powered).
  8. **Sub_metering_2**: energy sub-metering No. 2 (in watt-hour of active energy). It corresponds to the laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light.
  9. **Sub_metering_3**: energy sub-metering No. 3 (in watt-hour of active energy). It corresponds to an electric water-heater and an air-conditioner.

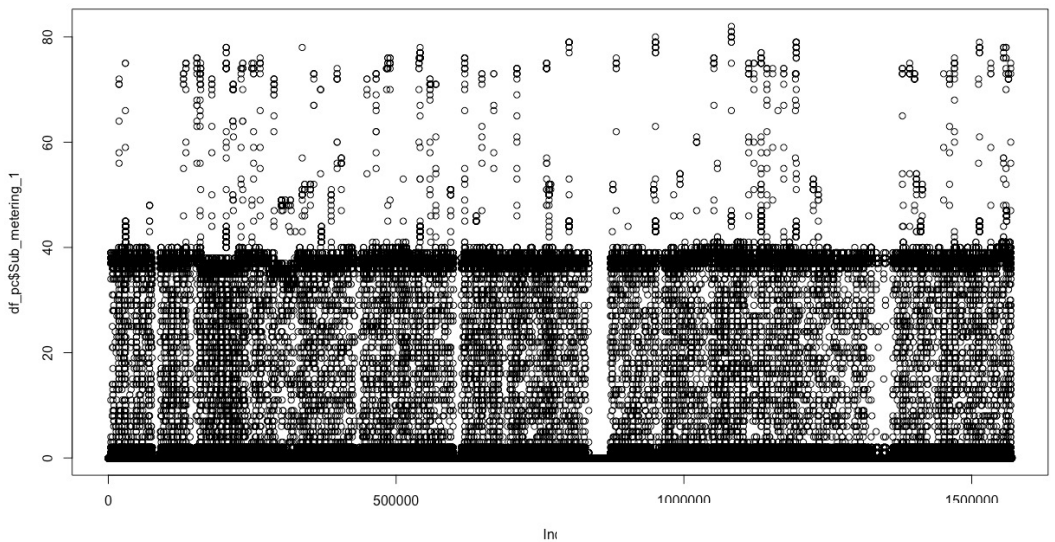
yr_2006	21992 obs. of 5 variables
yr_2007	521669 obs. of 5 variables
yr_2008	526905 obs. of 5 variables
yr_2009	521320 obs. of 5 variables
yr_2010	457394 obs. of 5 variables

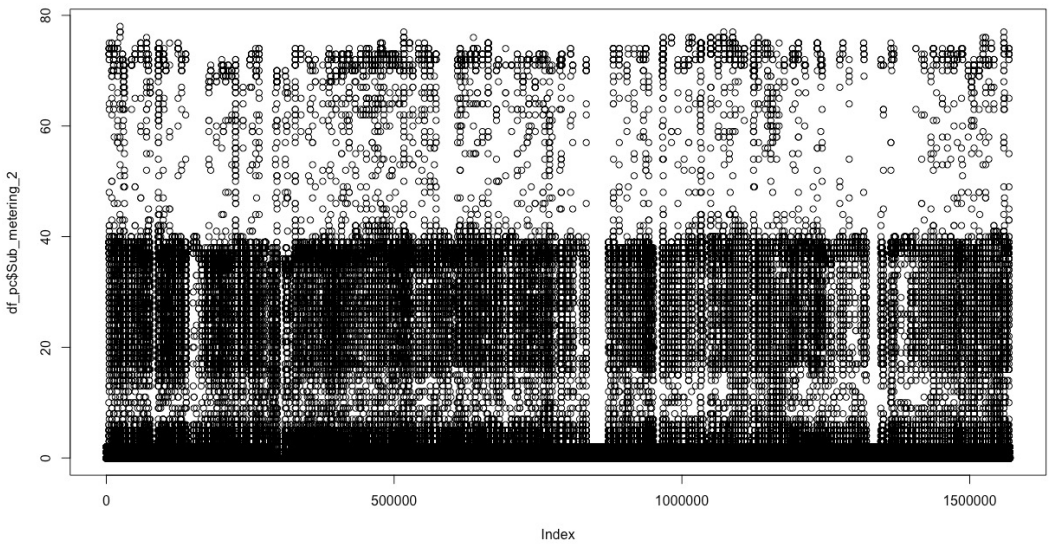
Initial Analysis and Findings

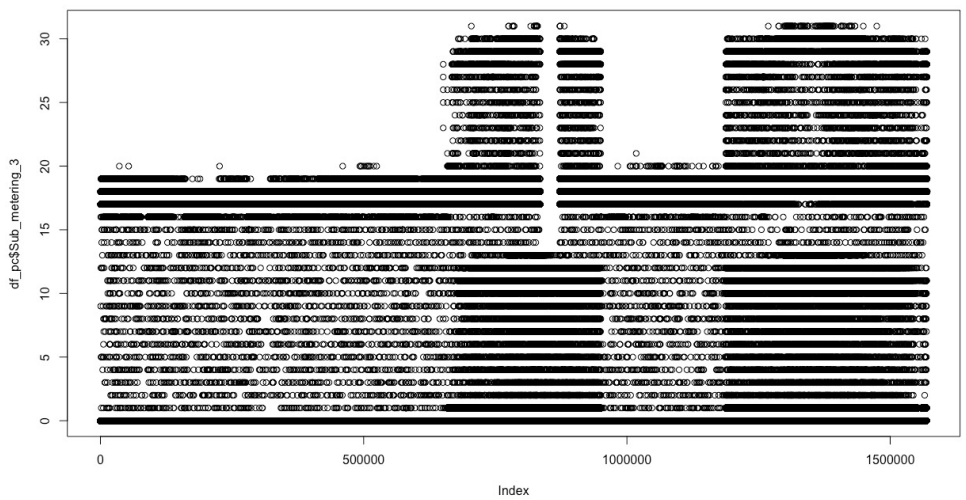
- Selected 2007,2008, and 2009 data for analysis as these are complete data for the year. 2006 and 2010 seems incomplete.
- The data said it was a house in Paris but when we converted the date and time, we found it was 1 hr. off so we used GMT.
- With mean of 6.22 we found sub meter 3 is mostly used which is the water heater and AC unit.
- The least used power is sub meter 1 which is the Kitchen.
- Even though sub meter has high 'mean' its maximum power consumption is less than sub meter 1. This means that one of the appliance in Kitchen is consuming more power than a large appliance like water heater and AC.
- With standard deviation we can tell sub meter 2 is evenly distributed than sub meter 1 and 3.

```
> summary(df_pc)
      year          month             day              hour          minute          week       Sub_metering_1   Sub_metering_2   Sub_metering_3
 Min.   :2007   Min.   : 1.000   Min.   : 1.00   Min.   : 0.0   Min.   : 0.00   Min.   : 1.00   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.:2007   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 6.0   1st Qu.:14.25   1st Qu.:13.00   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
 Median :2008   Median : 7.000   Median :16.00   Median :12.0   Median :30.00   Median :27.00   Median : 0.000   Median : 0.000   Median : 1.000
 Mean   :2008   Mean   : 6.528   Mean   :15.71   Mean   :11.5   Mean   :29.50   Mean   :26.62   Mean   : 1.159   Mean   : 1.343   Mean   : 6.216
 3rd Qu.:2009   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:18.0   3rd Qu.:44.00   3rd Qu.:40.00   3rd Qu.: 0.000   3rd Qu.: 1.000   3rd Qu.:17.000
 Max.   :2009   Max.   :12.000   Max.   :31.00   Max.   :23.0   Max.   :59.00   Max.   :53.00   Max.   :82.000   Max.   :78.000   Max.   :31.000
> psych::describe(df_pc)
               vars       n    mean    sd  min  max range   se
year              1 1569894 2008.00  0.82 2007 2009     2 0.00
month             2 1569894    6.53  3.46    1   12    11 0.00
day               3 1569894   15.71  8.80    1   31    30 0.01
hour              4 1569894   11.50  6.92    0   23    23 0.01
minute            5 1569894   29.50 17.32    0   59    59 0.01
week              6 1569894   26.62 15.10    1   53    52 0.01
Sub_metering_1    7 1569894    1.16  6.29    0   82    82 0.01
Sub_metering_2    8 1569894    1.34  5.97    0   78    78 0.00
Sub_metering_3    9 1569894    6.22  8.34    0   31    31 0.01
```

Initial Analysis - Plots



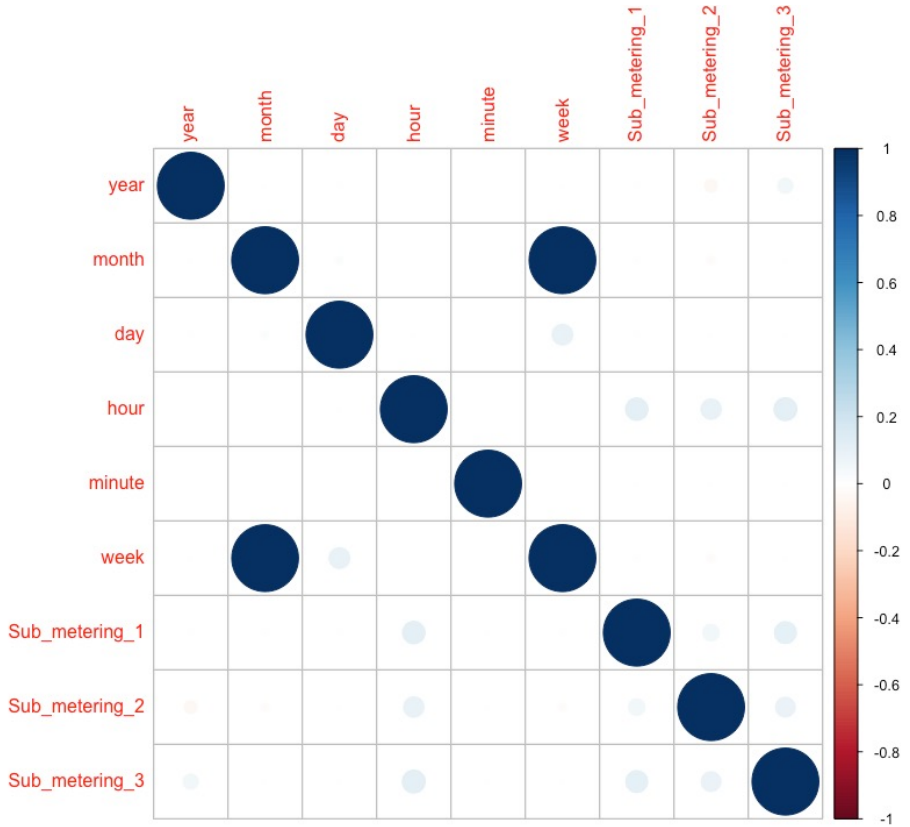


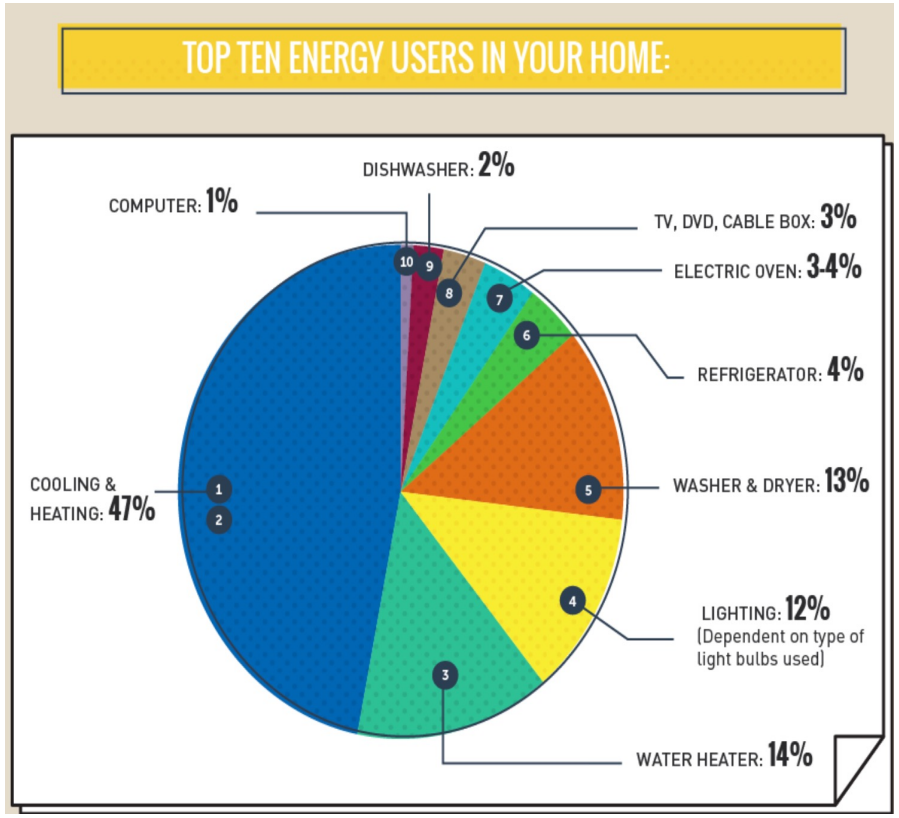


These plots show the sub meter 1 , 2 , and 3 distribution.
Notice significant gaps in sub meter 1 and 2. Sub meter 3 has minimum recording all times.

Initial Analysis - Correlation

- Sub meter 1 has positive correlation with hour, sub meter 2 and 3.
- Sub meter 2 is negatively correlated to year, positive with hour, sub meter 1 and 3.
- Sub meter 3 is positive with year, hour, sub meter 1 and 2.
- Hour and year are important to sub meters.
- Week and Month are highly correlated, and week is positive with day.





Recommendations

- Is there additional power and power related information that would benefit analytics in the future if added to the data set?
  - ✓ After few initial observations, we would recommend having the entire house data including other rooms power consumption would help better understand the occupancy question.
  - ✓ It would also be helpful to understand how the appliances are operated and if any of them are connected to the internet like IOT so remote access is applicable.
- Are there any changes to the sub-metering data collection structure that would help future analytics?
  - ✓ We would like to get complete **two or more years data prior** to the summer of 2008 and **after** so we can be solid in our analysis and conclusions.
- If you could add more information to the data set, what kinds of attributes would you add? What would be important to understanding the power usage in this home?
  - ✓ Maybe data like thermostat reading and other rooms data would help understand things better regarding occupancy.
  - ✓ A TV or Media equipment data would be very useful for occupancy analysis.
  - ✓ How big is the residence information
  - ✓ The age of the premise
  - ✓ The number of people living in the house
  - ✓ The type, number and age of appliances
  - ✓ Do the premise have a pool
  - ✓ Season information like define summer months – Is it from Marth to October?
- Should the appliances on the sub-meters be grouped the way they are currently grouped? Could more information be gained if some were separated?
  - ✓ We found the refrigerator in the laundry room and would be more advisable if sub meter 1 i.e., the kitchen has the refrigerator.
  - ✓ It would have been easy to the analysis if water heater and air condition had their own sub meter.

Q&A