HEPC Analysis

Introduction

In this mini-data analysis, individual household electric power consumption Data Set is used to do exploratory data analyais and visualization. The data is downloaded from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/individual+household+electric+power+consumption). It gives measurements of electric power consumption in one household near Paris with a one-minute sampling rate over a period of almost 4 years. Different electrical quantities and some sub-metering values are available including total power usage, kitchen, laundry, air conditioning, and other (the rest which is infered from the total and the three available categories).

Preliminary Analysis

```
library(data.table)
library(vcdExtra)
```

```
## Loading required package: vcd
```

```
## Loading required package: grid
```

```
## Loading required package: gnm
```

```
library(extracat)
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:vcdExtra':
##
##     summarise
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(zoo)
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(forcats)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

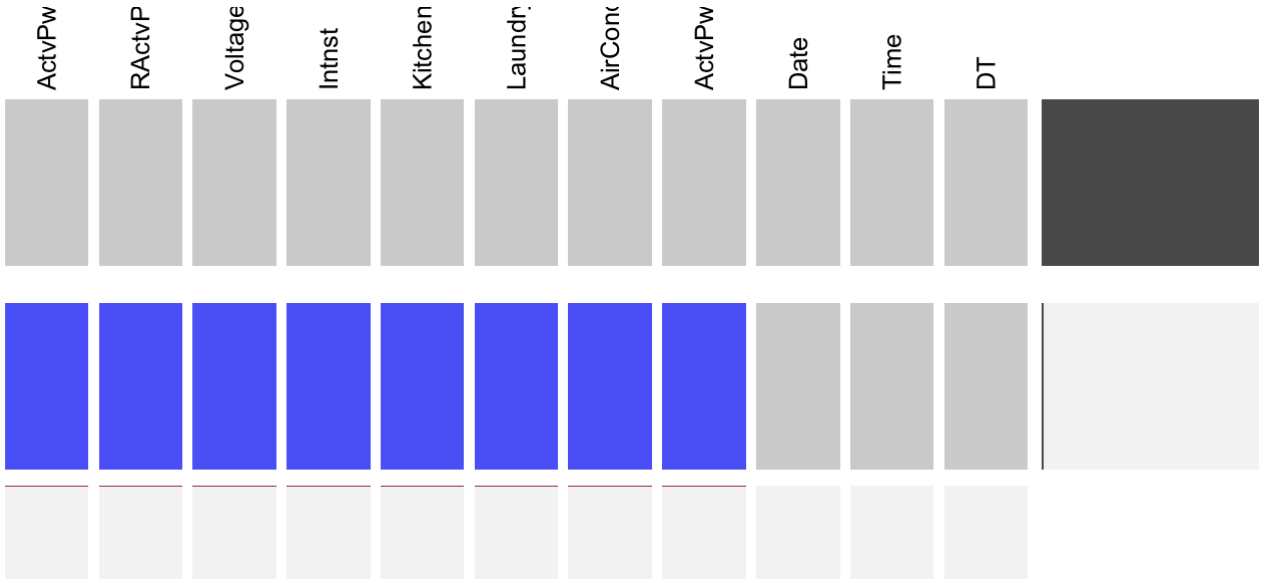
```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(ggplot2)
var_names <- c("Date", "Time", "ActvPwr", "RActvPwr","Voltage","Intnst","Kitchen", "Laun
dry", "AirCond")
df<-fread("household_power_consumption.txt",na.strings="?",col.names = var_names)
```

```
##
Read 74.2% of 2075259 rows
Read 2075259 rows and 9 (of 9) columns from 0.124 GB file in 00:00:03
```

```
df$Date <- as.Date(df$Date, format='%d/%m/%Y')
#df%>%mutate(DT=as.POSIXct(paste(Date,Time, sep = " "),"GMT"))->df
#df%>%mutate(DT=as.POSIXct(  do.call(paste, c(df[,1:2], sep = " ")), "GMT"))->df
#df%>%mutate(DT=paste(Date,Time, sep = " "))->df
#df%>%mutate(DateTime=as.POSIXct(DT,"Europe/Paris"))->df
df%>%mutate(DT=as.POSIXct(paste(Date,Time, sep = " "),"GMT"))->df
df%>%mutate(ActvPwr_wh=ActvPwr*1000/60)->df
```

```
visna(df,sort="b")
```




```
df%>%filter(is.na(ActvPwr))%>%dplyr::summarise(missing_pct=100.*n()/nrow(df))
```

```
##   missing_pct
## 1    1.251844
```

- This figure shows the missing pattern of the data. There is only one missing pattern with all data missing except the time stamps. The percentage of this missing pattern is 1.25% of the whole dataset. May looking into the Date and Time of the missing to see if they are randomly distrbuted.

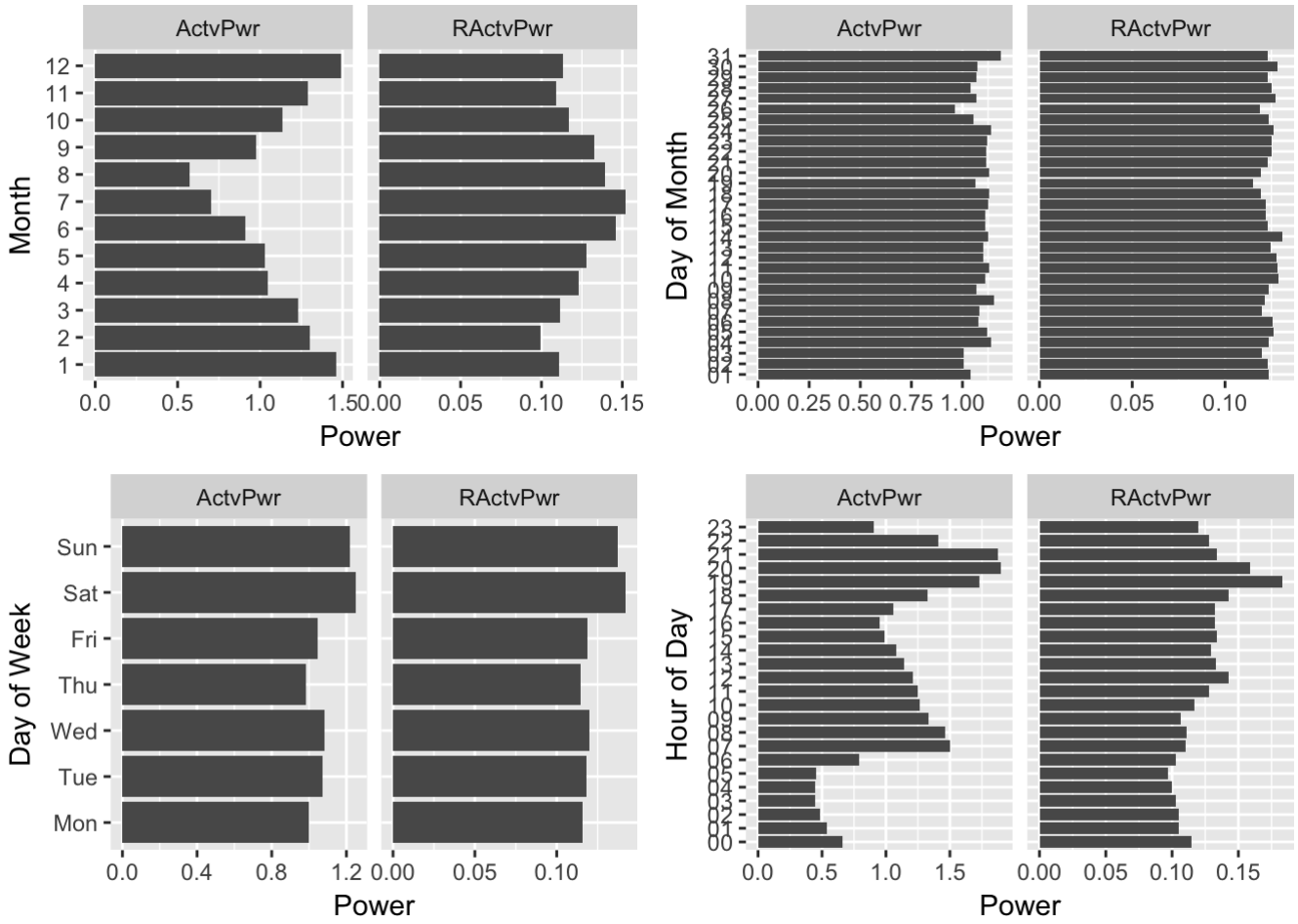
```
#df%>%select(DT,ActvPwr)%>%drop_na()%>%filter(DT<as.Date("2007-01-01 00:00:00 GMT"))%>%g
gplot()+
#  geom_line(aes(DT,ActvPwr))+
#  ggtitle("Minutely-household global minute-averaged active power")+
#  labs(x="Date",y="Kilowatt")+theme(legend.title=element_blank())
```

```
df%>%mutate(other=ActvPwr*1000/60-Kitchen-Laundry-AirCond)->df
```

```
df%>%select(Date,Time,ActvPwr,RActvPwr)%>%gather(key,value,-Date,-Time)->df1
df%>%select(Date,Time,Kitchen,Laundry,AirCond,other)%>%gather(key,value,-Date,-Time)->df
2
df%>%
  mutate(Kpct=Kitchen/ActvPwr_wh,Lpct=Laundry/ActvPwr_wh,
         Apct=AirCond/ActvPwr_wh,Opct=other/ActvPwr_wh)%>%
  select(Date,Time,Kpct,Lpct,Apct,Opct)%>%gather(key,value,-Date,-Time)->df3
```

```
#frequency by month
df1%>%drop_na()%>%mutate(Month=as.character(month(Date)))%>%group_by(Month,key)%>%dply
r::summarise(ave=mean(value))->byDatePwr_mon
byDatePwr_mon%>%ggplot(aes(fct_relevel(Month,"10","11","12",after=9),ave))+geom_bar(stat
="identity")+coord_flip()+ylab("Power")+facet_wrap(~key,scales="free_x")+xlab("Month")->
p1

#frequency by day
df1%>%drop_na()%>%mutate(Day=as.factor(format(Date,"%d")))%>%group_by(Day,key)%>%dplyr::
summarise(ave=mean(value))->byDatePwr_day
byDatePwr_day%>%ggplot(aes(Day,ave))+geom_bar(stat="identity")+coord_flip()+ylab("Power"
)+facet_wrap(~key,scales="free_x")+xlab("Day of Month")->p2

#frequency by weekday
df1%>%drop_na()%>%mutate(Wkday=as.factor(weekdays(Date,abbreviate=TRUE)))%>%group_by(Wkd
ay,key)%>%dplyr::summarise(ave=mean(value))->byDatePwr_wkday

byDatePwr_wkday%>%ggplot(aes(fct_relevel(Wkday,"Mon","Tue","Wed","Thu","Fri","Sat","Sun"
),ave))+geom_bar(stat="identity")+coord_flip()+ylab("Power")+facet_wrap(~key,scales="fre
e_x")+xlab("Day of Week")->p3

#Frequency by hour of day, combining hour 00 and hour 24 into hour 00
df1%>%drop_na()%>%mutate(Hour=as.factor(substr(Time,1,2)))%>%group_by(Hour,key)%>%dply
r::summarise(ave=mean(value))->byDatePwr_hour
byDatePwr_hour$Hour<-factor(byDatePwr_hour$Hour)

byDatePwr_hour%>%ggplot(aes(Hour,ave))+geom_bar(stat="identity")+coord_flip()+ylab("Powe
r")+facet_wrap(~key,scales="free_x")+xlab("Hour of Day")->p4
```

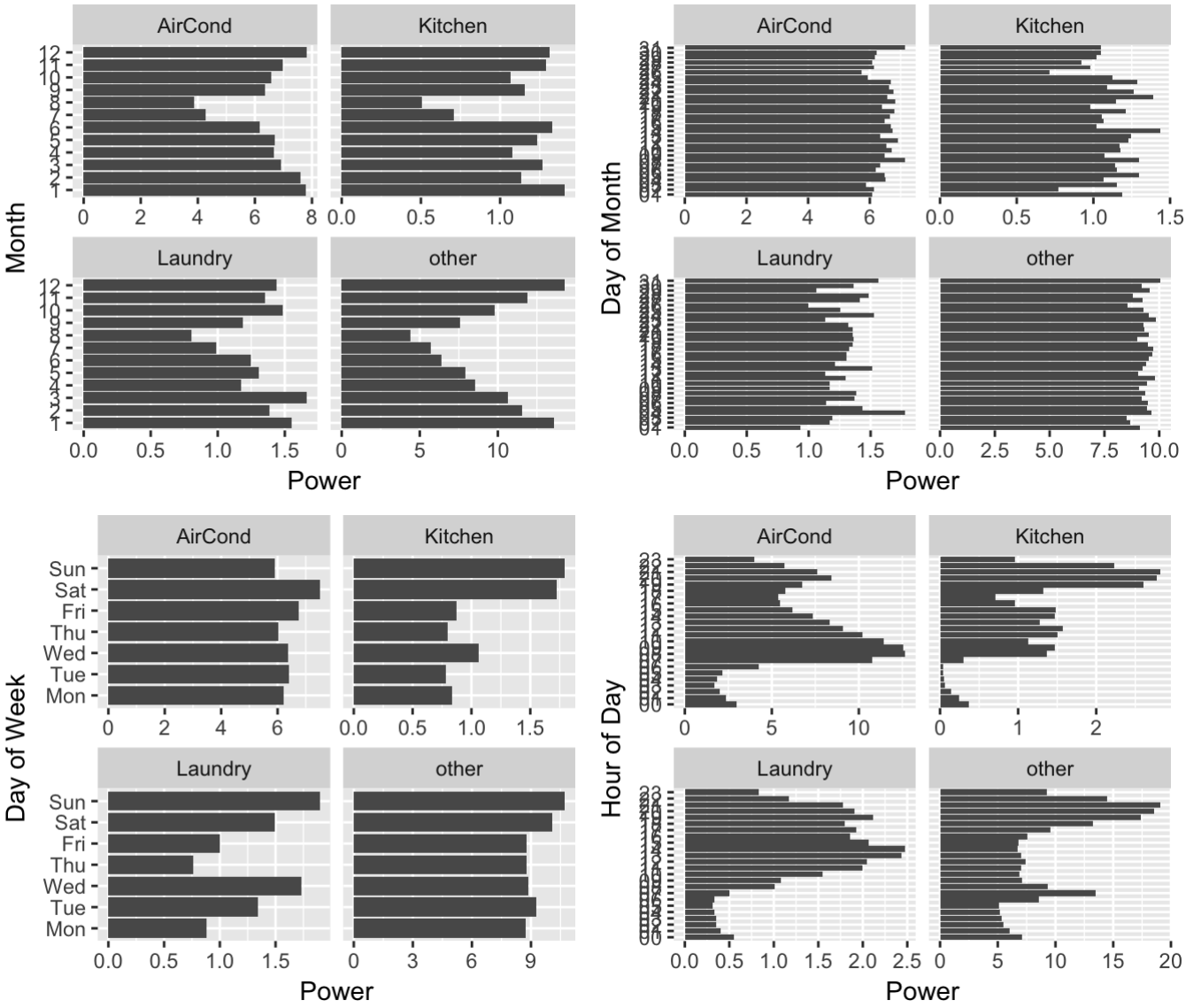
```
grid.arrange(p1,p2,p3,p4,nrow=2)
```



- This plots shows the anaual cycle, montly cycle, weekly cycle and daily cycle of the power usage pattern (ActvPwr is the actual power usage).
- July and August has the least usage among a year.
- Saturday and Sunday has the peak usage during a weekly.
- Morning and evening peaks in a daily cycle.
- Looks like a typical working family pattern with high usage during weekends and eveningtime, with low usage during weekdays and daytime.

```
#frequency by month
df2%>%drop_na()%>%mutate(Month=as.character(month(Date)))%>%group_by(Month,key)%>%dply
r::summarise(ave=mean(value))->byDatePwr_mon
byDatePwr_mon%>%ggplot(aes(fct_relevel(Month,"10","11","12",after=9),ave))+geom_bar(stat
="identity")+coord_flip()+ylab("Power")+facet_wrap(~key,scales="free_x")+xlab("Month")->
p1

#frequency by day
df2%>%drop_na()%>%mutate(Day=as.factor(format(Date,"%d")))%>%group_by(Day,key)%>%dplyr::
summarise(ave=mean(value))->byDatePwr_day
byDatePwr_day%>%ggplot(aes(Day,ave))+geom_bar(stat="identity")+coord_flip()+ylab("Power"
)+facet_wrap(~key,scales="free_x")+xlab("Day of Month")->p2

#frequency by weekday
df2%>%drop_na()%>%mutate(Wkday=as.factor(weekdays(Date,abbreviate=TRUE)))%>%group_by(Wkd
ay,key)%>%dplyr::summarise(ave=mean(value))->byDatePwr_wkday

byDatePwr_wkday%>%ggplot(aes(fct_relevel(Wkday,"Mon","Tue","Wed","Thu","Fri","Sat","Sun"
),ave))+geom_bar(stat="identity")+coord_flip()+ylab("Power")+facet_wrap(~key,scales="fre
e_x")+xlab("Day of Week")->p3

#Frequency by hour of day, combining hour 00 and hour 24 into hour 00
df2%>%drop_na()%>%mutate(Hour=as.factor(substr(Time,1,2)))%>%group_by(Hour,key)%>%dply
r::summarise(ave=mean(value))->byDatePwr_hour
byDatePwr_hour$Hour<-factor(byDatePwr_hour$Hour)

byDatePwr_hour%>%ggplot(aes(Hour,ave))+geom_bar(stat="identity")+coord_flip()+ylab("Powe
r")+facet_wrap(~key,scales="free_x")+xlab("Hour of Day")->p4
```

```
grid.arrange(p1,p2,p3,p4,nrow=2)
```



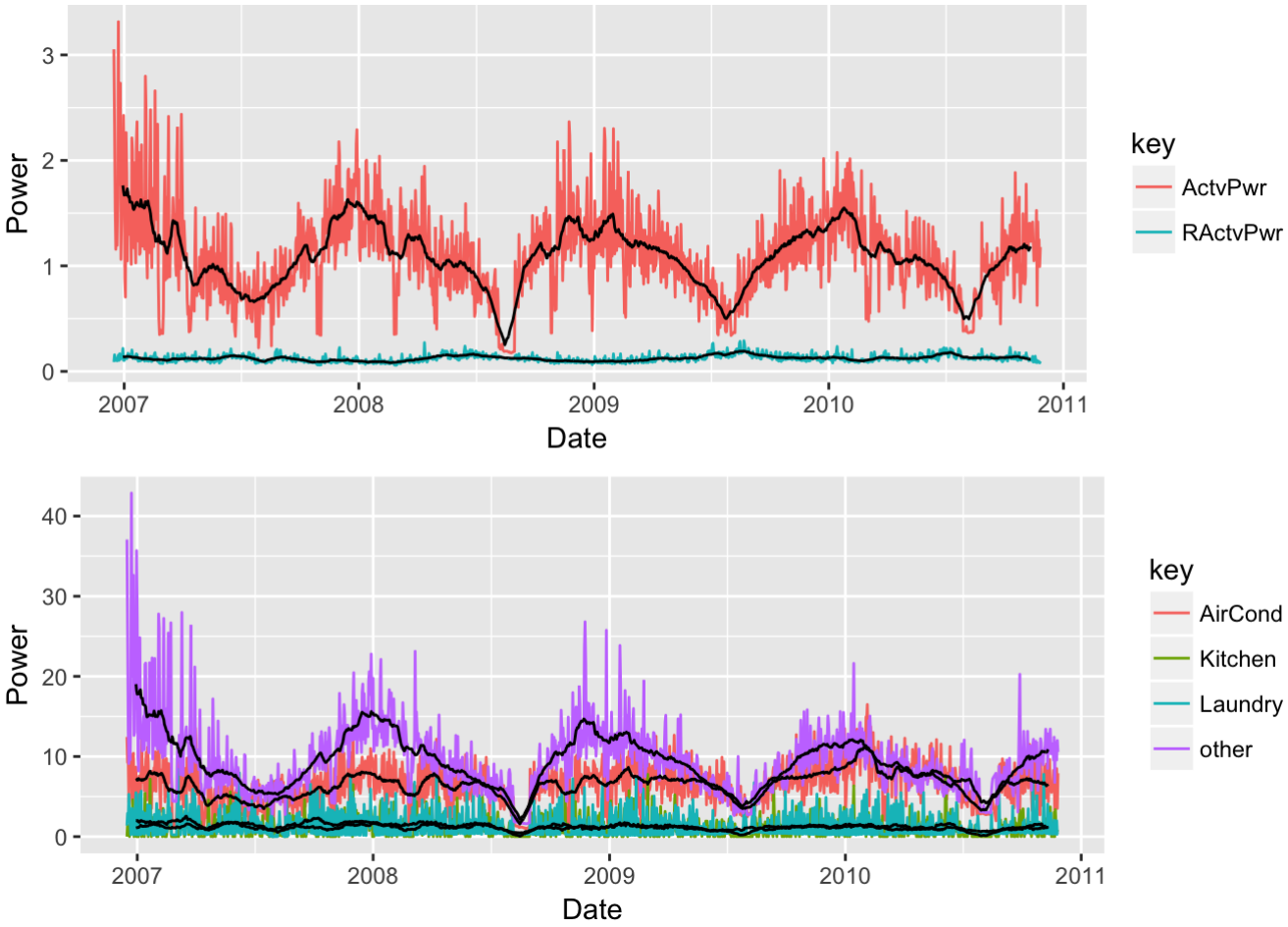
- This figure looks at sub-metering data for the kitchen, air-conditioning+ water heater, Laundry, and other (the rest).
- The low usage during summer time is contributed from all of the 4 categories.
- Laundry usage peaks at Wednesday and Sunday, twice a week.
- Kitchen usage is highest during weekend. Saturday also shows a peak in air-conditioning+water heater. For the other category which excludes kitchen, laundry, and air-conditioning also has a peak during weekend.
- In the daily cycle, kitchen usage is peaked during evening hours while air conditioning usage peaked at morning hours.Laundry peaked after noon and later evening. The other category also has a peak in the late evening and a second peak in the morning.

```
#daily time series
df1%>%drop_na()%>%group_by(Date,key)%>%dplyr::summarise(ave=mean(value))%>%ungroup()%>%g
roup_by(key)%>%mutate(mon_mean=rollmean(ave,30,fill=NA))%>%ungroup()->byDatePwr_day
byDatePwr_day%>%ggplot()+geom_line(aes(Date,ave,color=key))+geom_line(aes(Date,mon_mean,
group=key))+ylab("Power")+xlab("Date")->p1
```

```
#daily time series
df2%>%drop_na()%>%group_by(Date,key)%>%dplyr::summarise(ave=mean(value))%>%ungroup()%>%g
roup_by(key)%>%mutate(mon_mean=rollmean(ave,30,fill=NA))%>%ungroup()->byDatePwr_day
byDatePwr_day%>%ggplot()+geom_line(aes(Date,ave,color=key))+geom_line(aes(Date,mon_mean,
group=key))+ylab("Power")+xlab("Date")->p2
grid.arrange(p1,p2,nrow=2)
```

```
## Warning: Removed 58 rows containing missing values (geom_path).
```

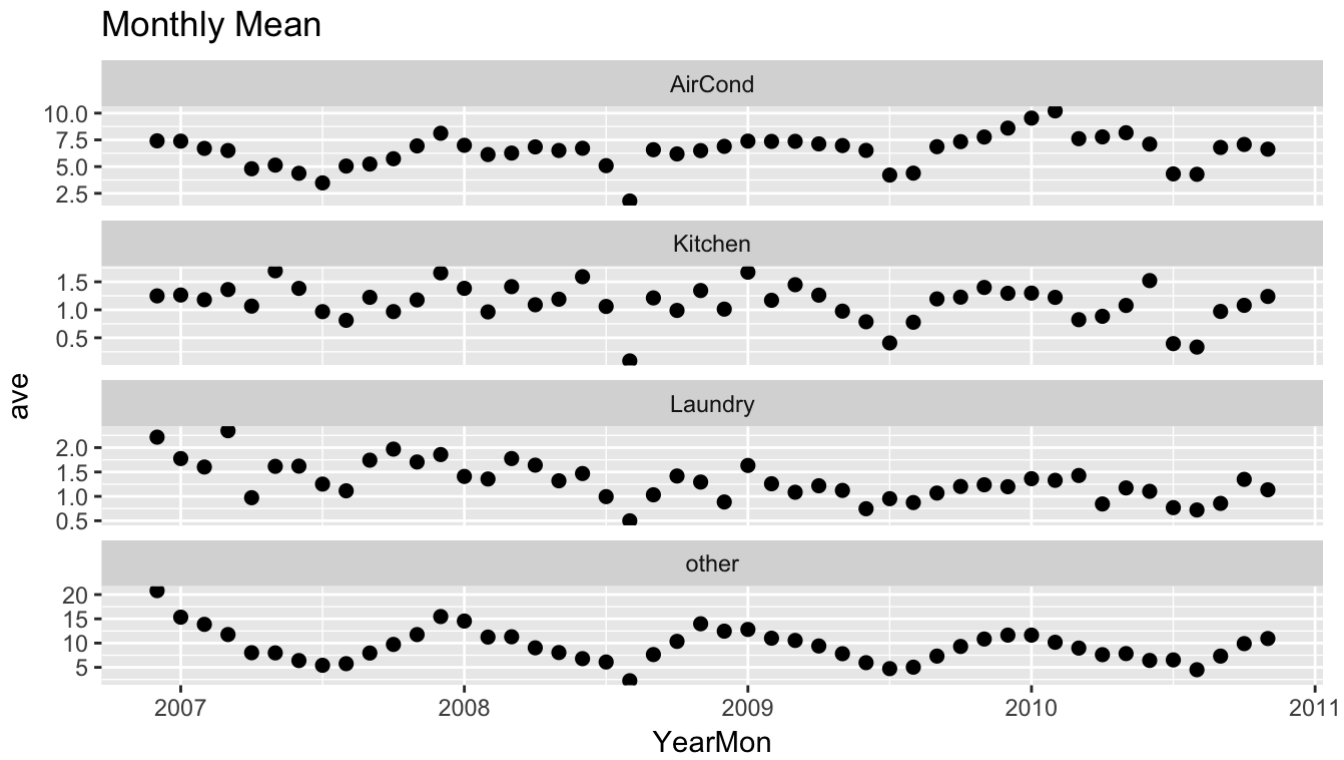
```
## Warning: Removed 116 rows containing missing values (geom_path).
```



- Daily time series shows annual cycle with a minimum during summer.

```
df2%>%drop_na()%>%mutate(YearMon =as.yearmon(Date))%>%group_by(YearMon,key)%>%dplyr::sum
marise(ave=mean(value))->meanByMonth

ggplot(meanByMonth)+geom_point(aes(YearMon,ave),size=2)+ggtitle("Monthly Mean")+facet_wr
ap(~key,scales="free_y",nrow=4)
```

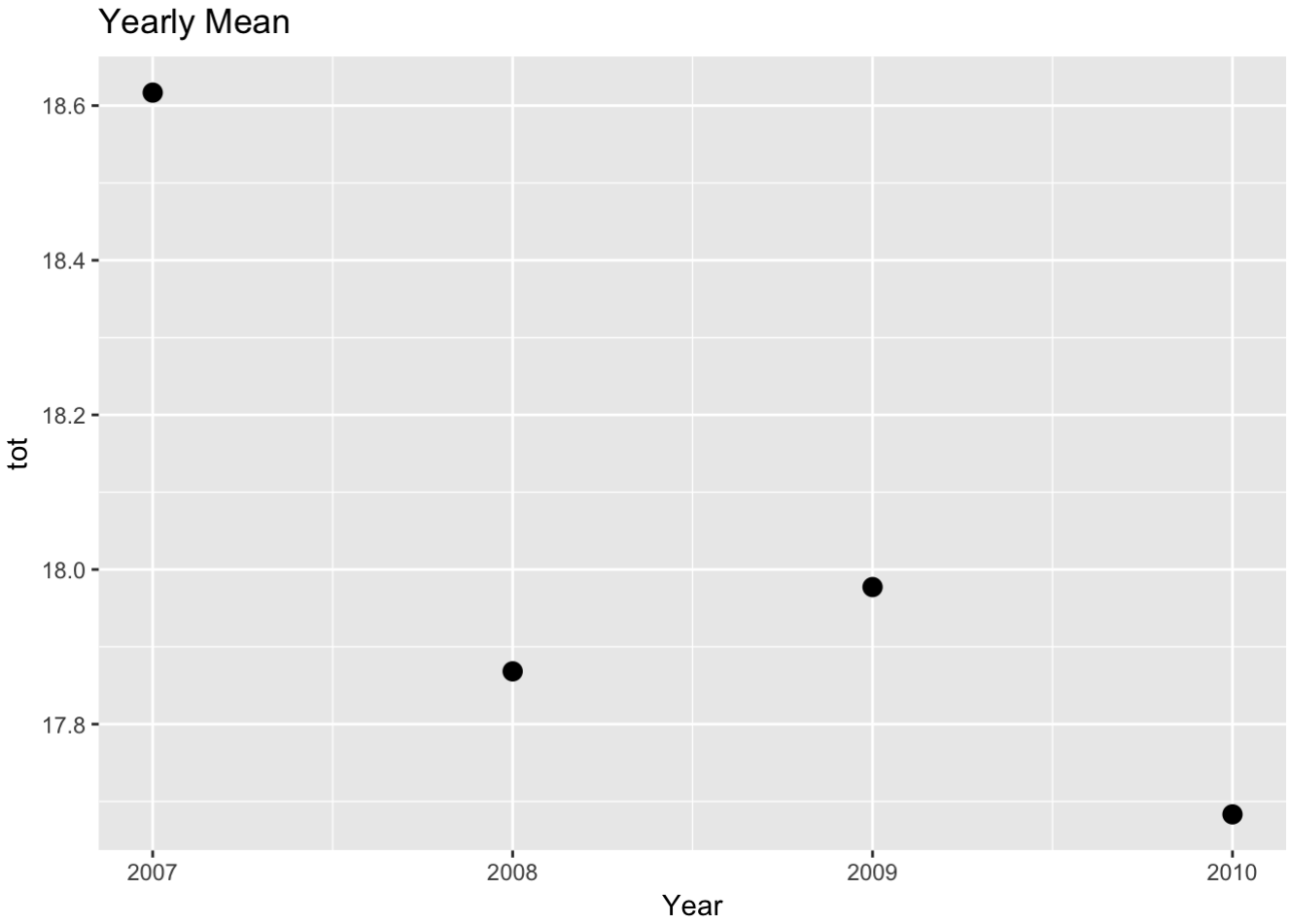
```
## Don't know how to automatically pick scale for object of type yearmon. Defaulting to
continuous.
```



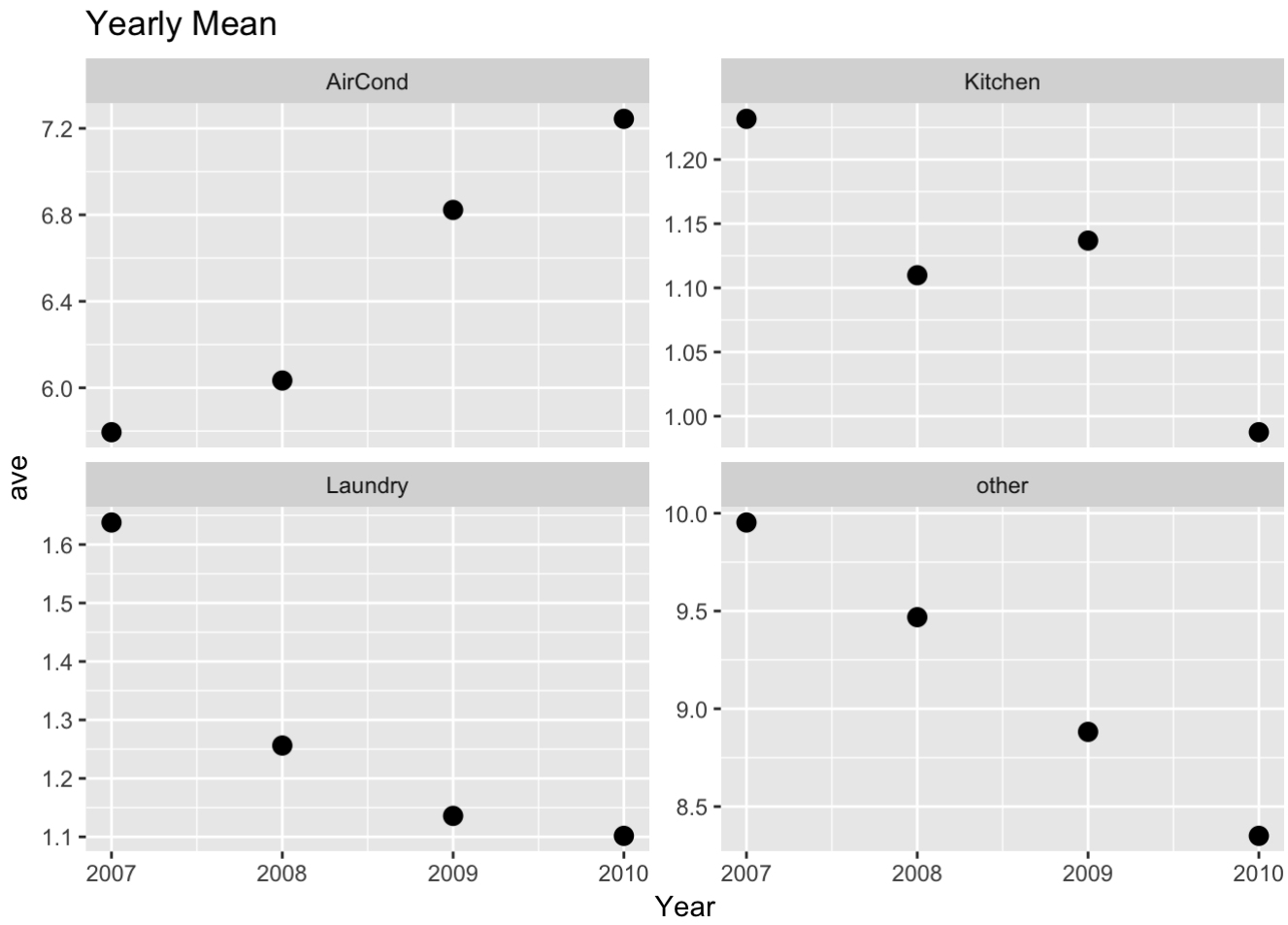
- Monthly mean time series show similar pattern as the daily. The majority of the power usage are from air-conditioning & water heater and from the other category.

```
df2%>%drop_na()%>%mutate(Year =year(Date))%>%group_by(Year,key)%>%dplyr::summarise(ave=m
ean(value))->meanByYear
df3%>%drop_na()%>%mutate(Year =year(Date))%>%group_by(Year,key)%>%dplyr::summarise(ave_p
ct=100*mean(value))->meanByYear_pct
```

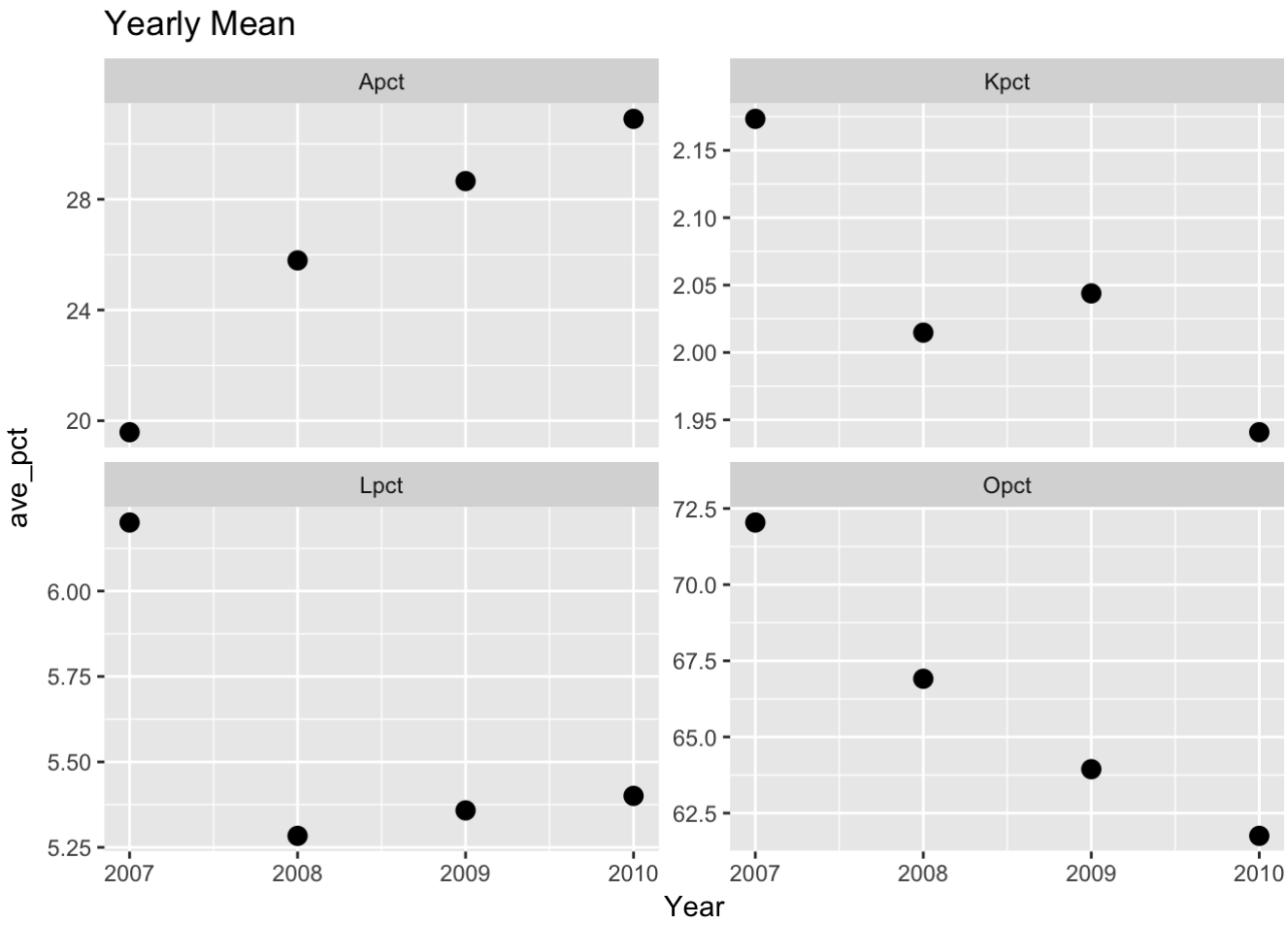
```
meanByYear%>%filter(Year>=2007)%>%group_by(Year)%>%dplyr::summarise(tot=sum(ave))%>%ggpl
ot()+geom_point(aes(Year,tot),size=3)+ggtitle("Yearly Mean")+xlim(2007,2010)
```




```
meanByYear%>%filter(Year>=2007)%>%ggplot()+geom_point(aes(Year,ave),size=3)+ggtitle("Yea
rly Mean")+facet_wrap(~key,scales="free_y",nrow=2)+xlim(2007,2010)
```




```
meanByYear_pct%>%filter(Year>=2007)%>%ggplot()+geom_point(aes(Year,ave_pct),size=3)+ggti
tle("Yearly Mean")+facet_wrap(~key,scales="free_y",nrow=2)+xlim(2007,2010)
```



- Year 2007 has the highest mean yearly usage. The rest of the 3 years some what similar.
- The usage of air conditioning+water heater increased over the year.
- The usage of the other, the kitchen and the laundry shows a rough decreasing over the 4 years
- The percentage usage are roughly the same as the absolute pattern
- With the dominant categories of air-conditioninig+water heater and the other compensating each other, the overall yearly mean decreased.
- The increasing usage of air-conditioninig and the decreasing usage in the other category needs further exploration to explain. Maybe due to the different weather.