

ALY6020 Final Poroject

Time Series for Individual household electric power consumption



Instructor:Dr. Justin Rodger

Northeastern University

December 10, 2018

Group Member: Shengyuan He, Shuyao Liang,

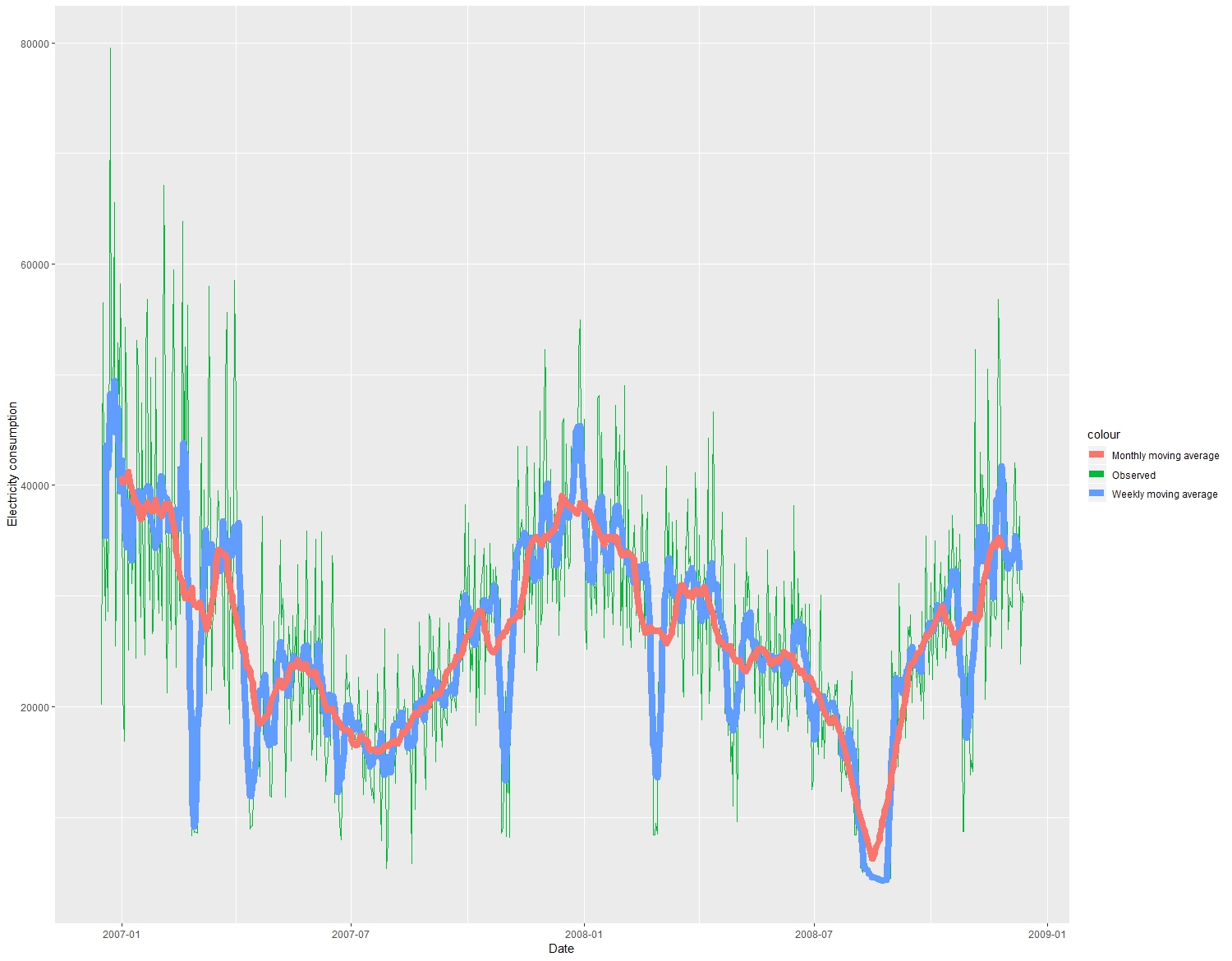
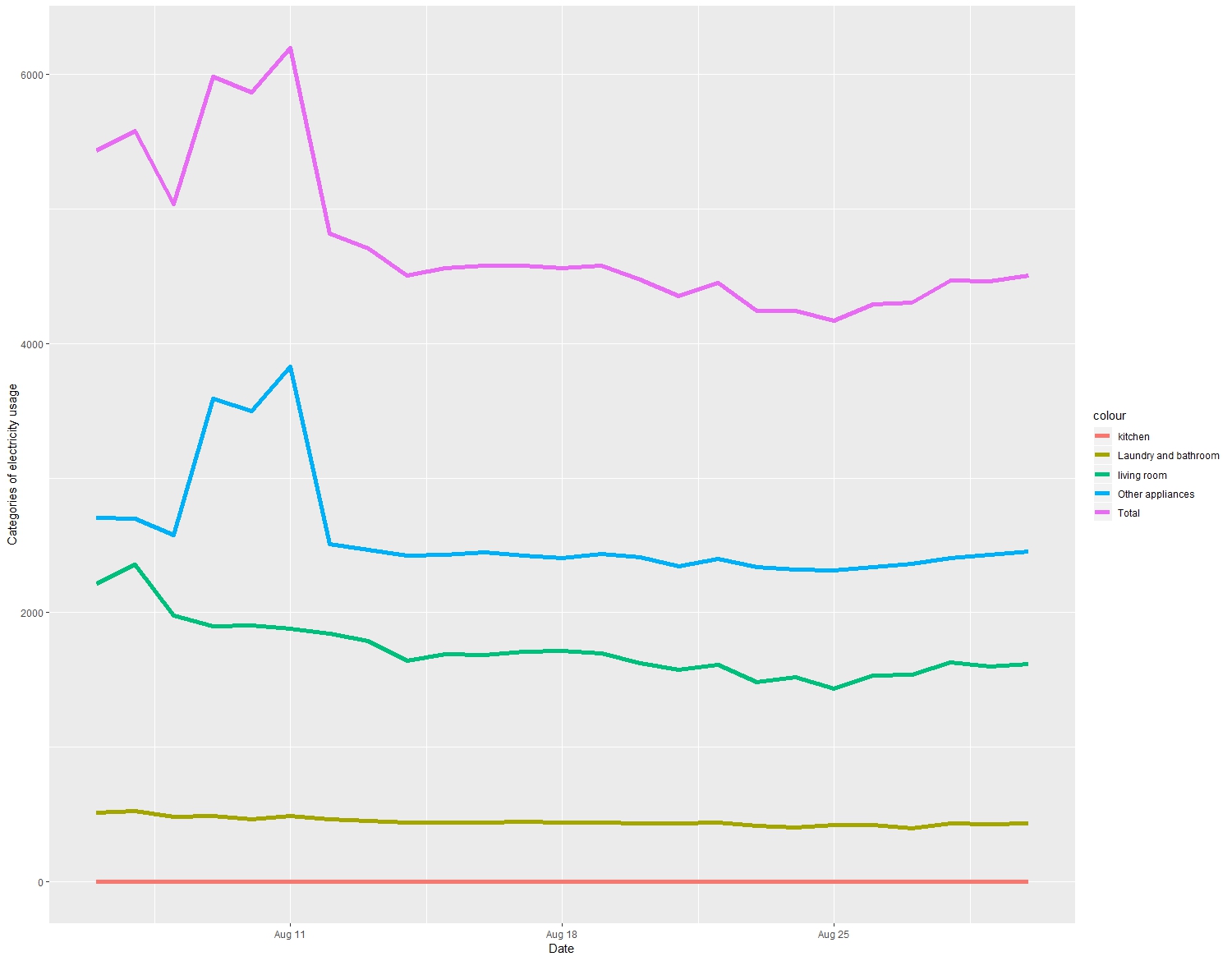
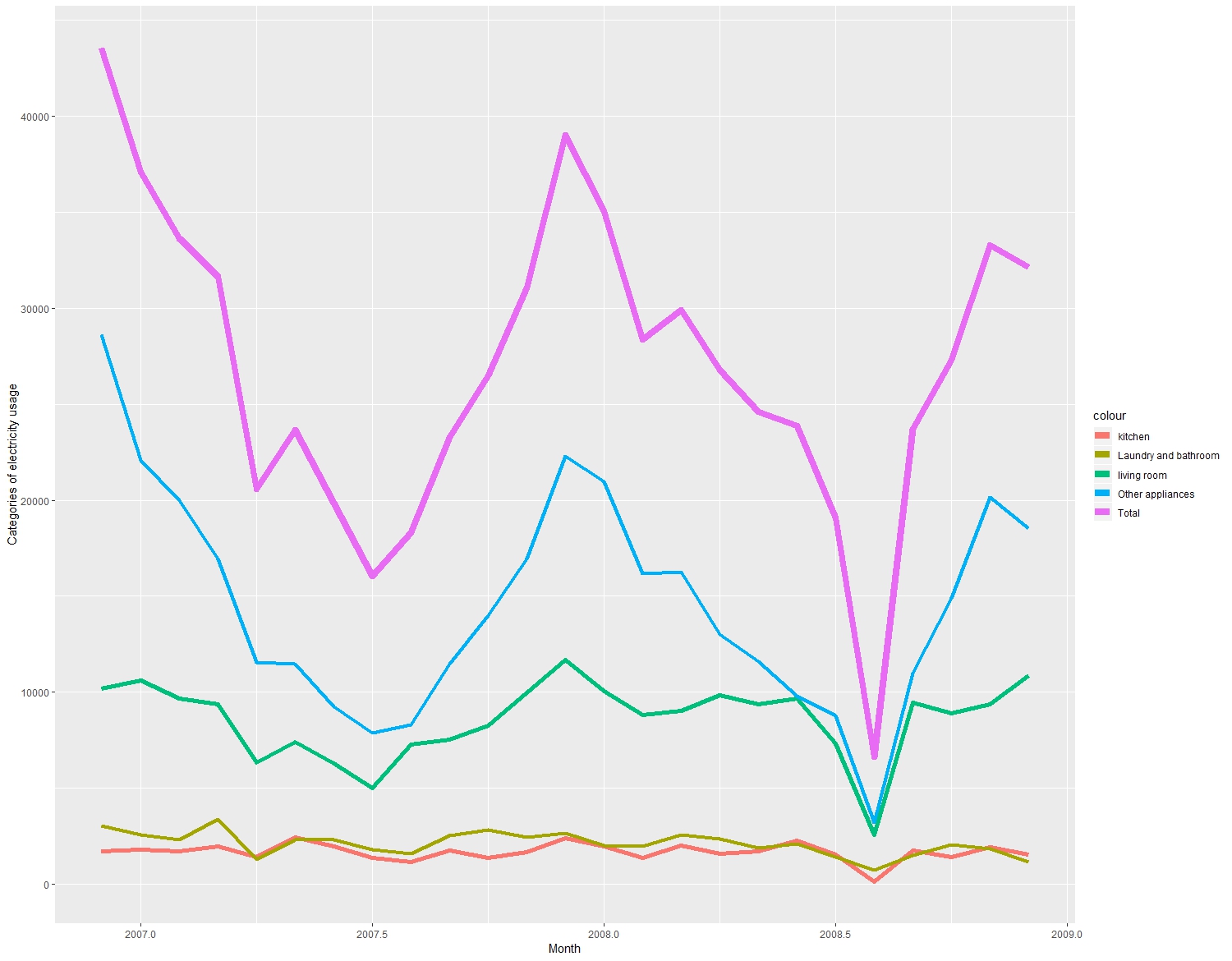
Songyuan Luo, Xinying Han

**Introduction-** Background and exploratory analysis

Nowadays, lots of businesses are adopting information technology to provide up-to-date information and use the information to maximize the benefits and methods by providing enough storage units to assist them to work. Planning by predicting future trends is a method to apply statistical knowledge to analyze data. In other words, the past is related to the current event. The result is used to predict future events. Time series is sequential historical data, similar to groups or observations. Data collected based on time is continuous for a period of time. The data might already exist in daily, weekly, monthly quarterly or yearly format, depending on which one is suitable for use. There are four parts of the time series: seasonal effects, trends, periodic and irregular effects. The time series analysis is using predictive techniques to identify past data from the model. People could assume that the information will be similar to people in the future events from the data that occurred. In this project, we will do the time series analysis with the Box and Jenkins method. We choose the Autoregressive Integrated Moving Average (ARIMA) and applied this method for discovering patterns and trends of the electric power in monthly. The most suitable forecasting method and the best choice of period is the value of Akaike Information Criterion (AIC).

The dataset we choose is ‘Individual household electric power consumption data set’. The data set contains 1048575 measurements gathered in a house located in Sceaux between December 2006 and November 2010. The total month is 47 months. Sceaux is 7 kilometers away from Paris, France. This dataset contains some missing value; the missing value is nearly 1.25% of the rows.

Data Visualization

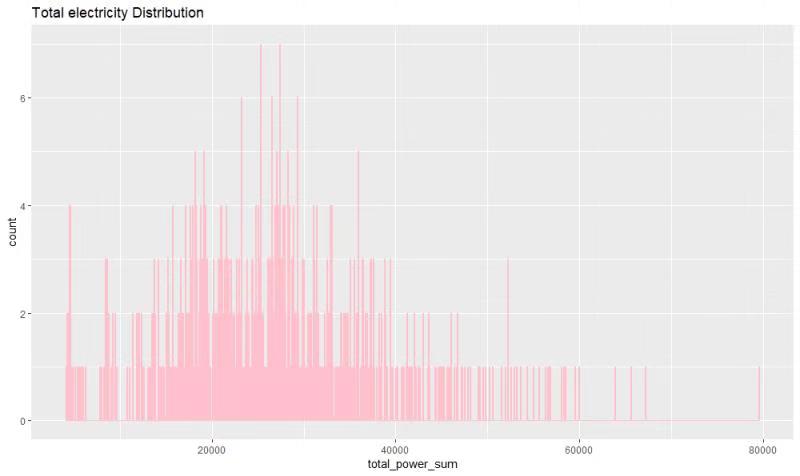
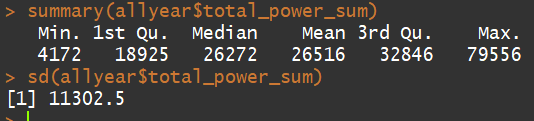


Adopted data visualization technology to visualize the categories usage in one chart. Expected total usage, category “other applications” always possess the highest position in the chart. Another chart is about moving average, which gives the trend line of electricity usage monthly and weekly from year 2007 to 2009.

**Problems and solution**

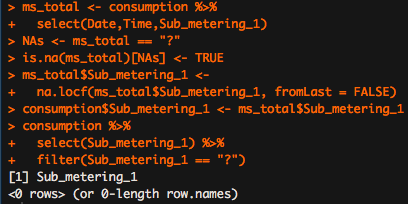
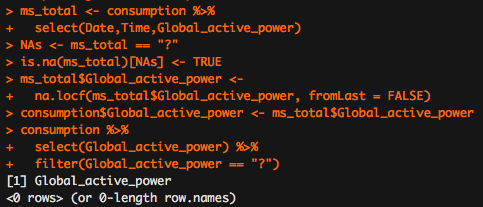
The purpose of this project is to find a predict model to forecast the electricity consumption in a household and to find the most suitable forecasting period whether it should be in daily or monthly. The solution is adopting forecasting techniques: Autoregressive Integrated Moving Average (ARIMA). Applied this method for detecting patterns and trends of the electric power consumption in the household with real time series period on daily and monthly. The most suitable forecasting method and the best choice of period were chosen by considering the smallest value of AIC.

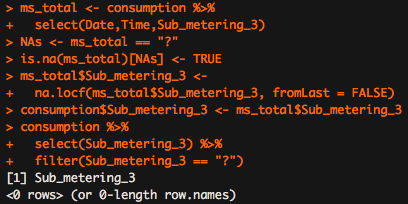
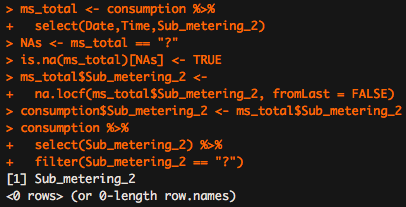
**Data Preparation-** Data cleaning

****

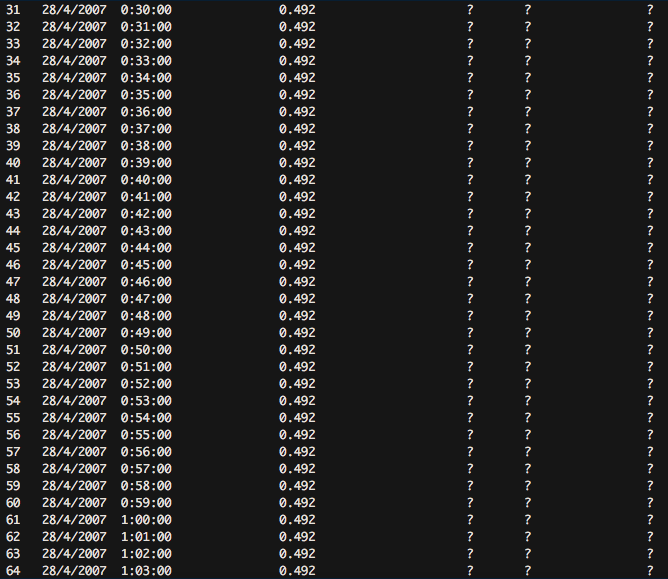
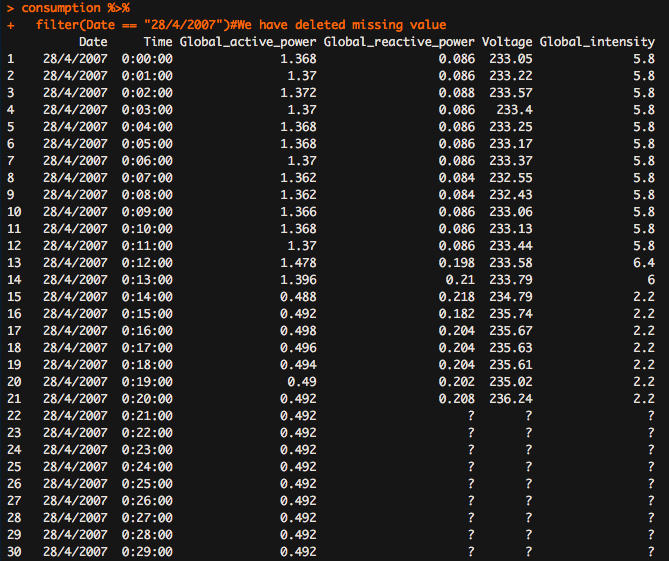
Total electricity distribution chart shows that the highest total\_power\_sum is around 30,000 kilowatt, and the intensity of section [20,000, 40,000] is the highest, which made a right skewed. Summary of the statistics presents the mean, standard deviation value of the total\_power\_sum.

According to UCI description, there are some missing value, consists of 1.25% of total dataset, and consecutive semicolons mean missing value. In this step, we will replace semicolon value by using previous value. For example, replace the ‘global\_active\_power’. The missing value may decrease the predictive efficiency of the forecasting model. We will fill missing data with the assumption that the current data will be similar to previous data. We will replace three columns: ‘sub\_meter\_1’, ‘sub\_meter\_2’ and ‘sum\_meter\_3’.

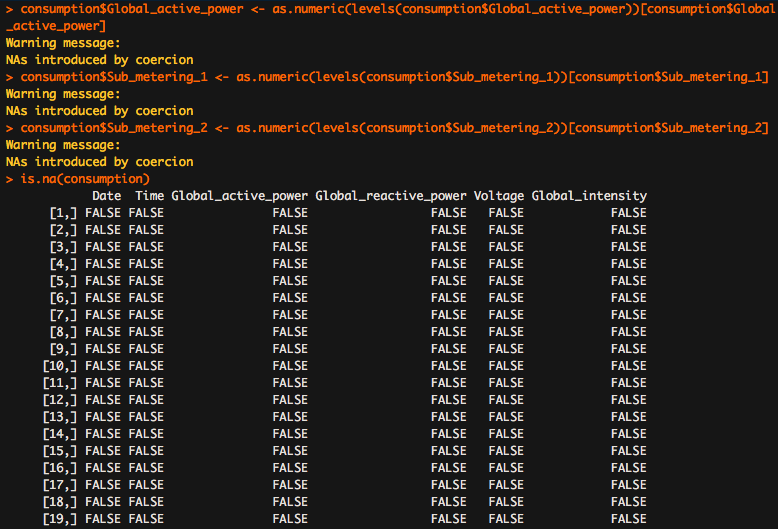




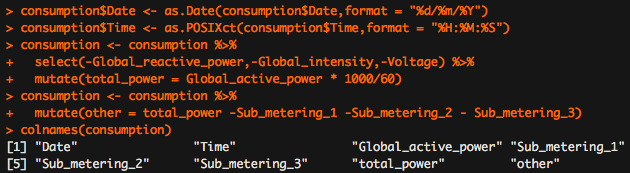
The next step is to test semicolon value. We could see there are “?” in the columns. We need to replace the “?”. We have deleted missing value before.



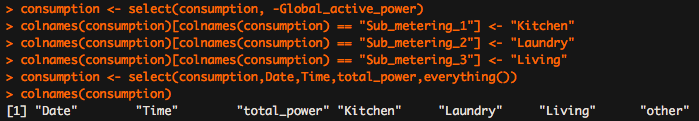
Then, we add new columns and edit column names for clearance. In this step, we pay attention on active energy consumption, which means we will be deleting Voltage, reactive power, intensity. And we will add other appliance column for reference. We transformed the factor to numeric. We could see the head of the new dataset.



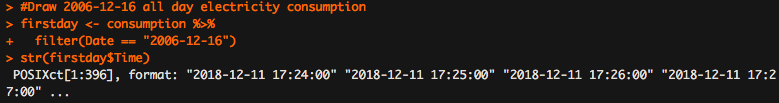
Add two new columns, one is the ‘total\_power’ and the other one is ‘other’.



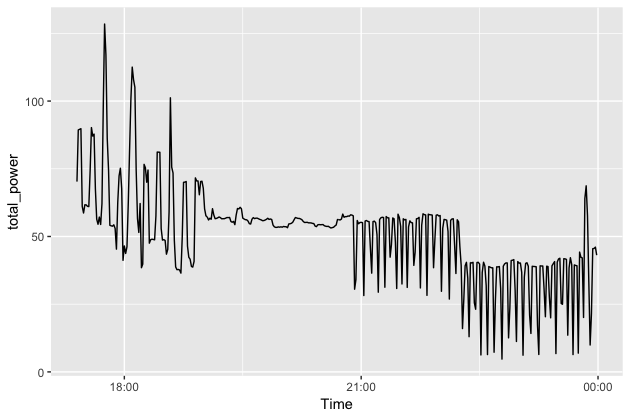
The next we checked that if the column names are changed successfully.



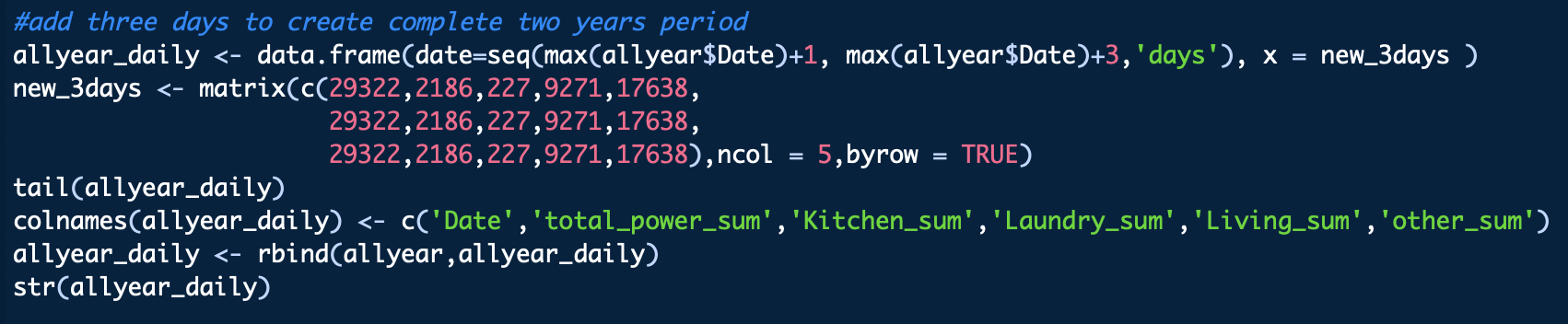
This step is to check the column names are changed successfully.

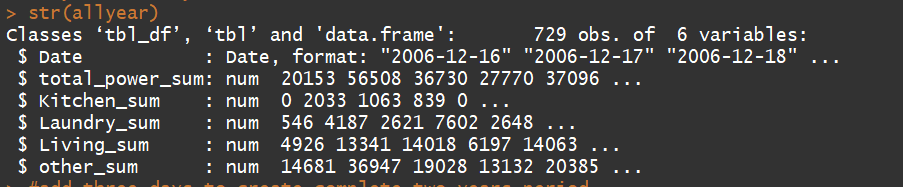
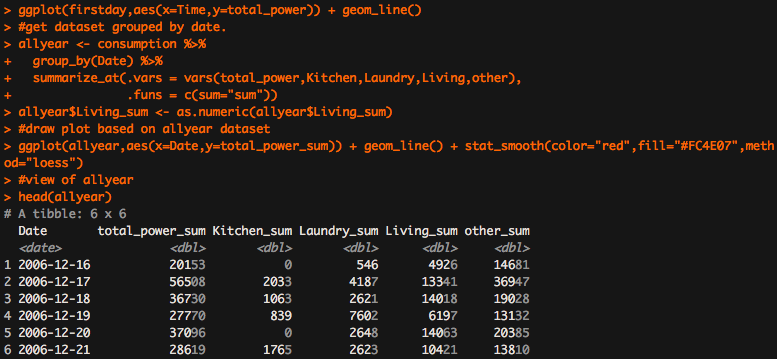


This graph is about the electricity usage of first day.

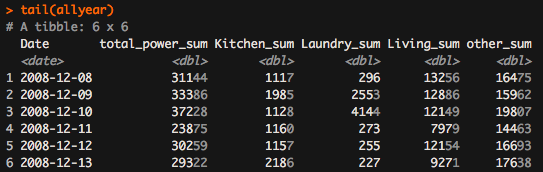
****

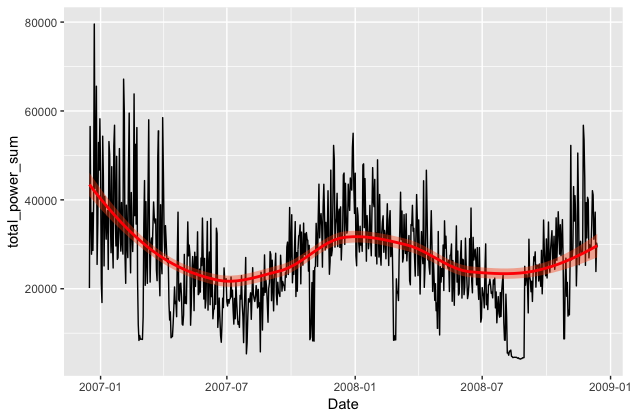
The next step is to check the electricity usage of all year. We could also see the ‘str’ of the dataset.



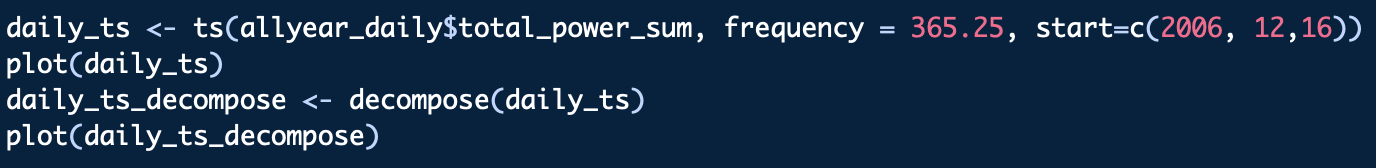


We used the ‘group’ command to collect the all dates and then draw the plot based on ‘allyear’ dataset. People also could see the trend-line.

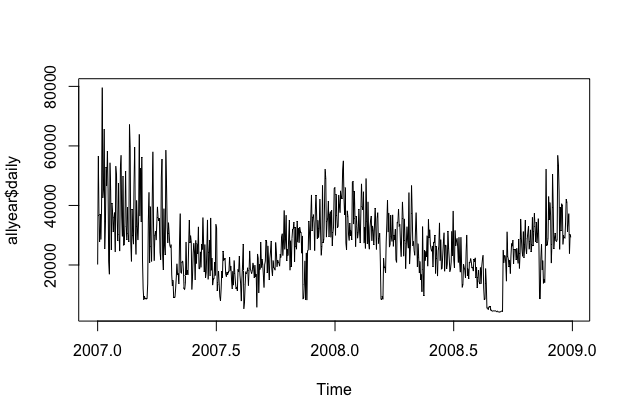




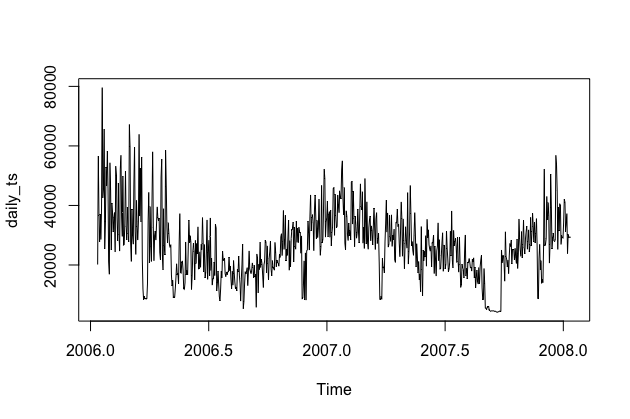
This is the daily analysis. We could see the trend of daily is the picture below.



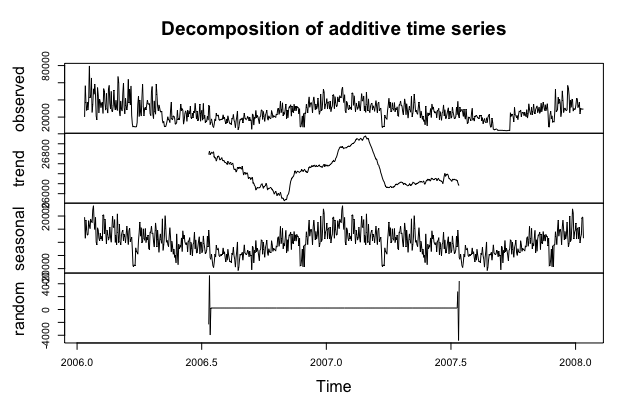
This is the daily plot. We could see the trend of daily analysis.



The next is about the time series model of daily.



The next graph is the Decomposition of additive time series.

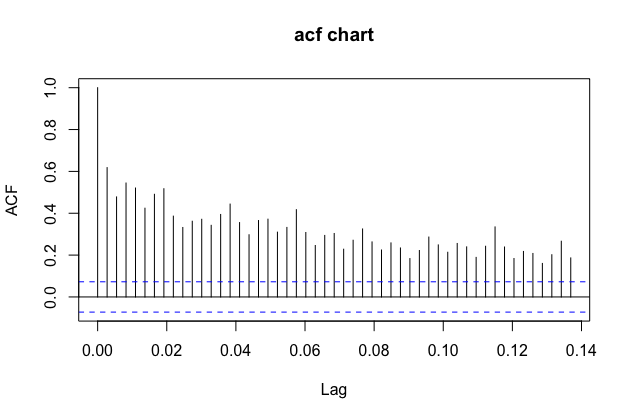
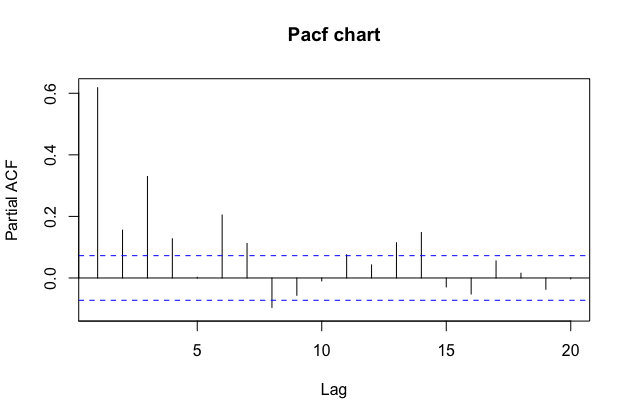


**Analysis**

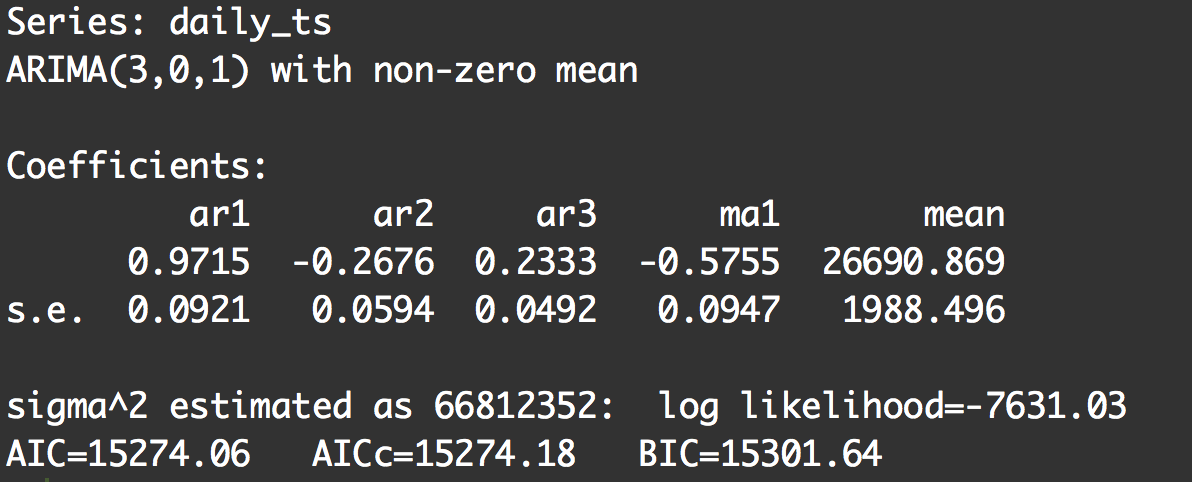
**Daily Time Series Model Analysis**

Daily Time Series Model - ACF and PACF

From ACF and PACF, it is very hard to determine the parameters. The PACF shown in Figure is suggestive of an AR(3) model. So an initial candidate model is an ARIMA (3,1,0). There are no other obvious candidate models. We fit an ARIMA (3,0,0) model along with variations including ARIMA(4,0,0), ARIMA(2,0,0), ARIMA(3,0,1), etc. Of these, the ARIMA (3,0,1) has a slightly smaller AICc value

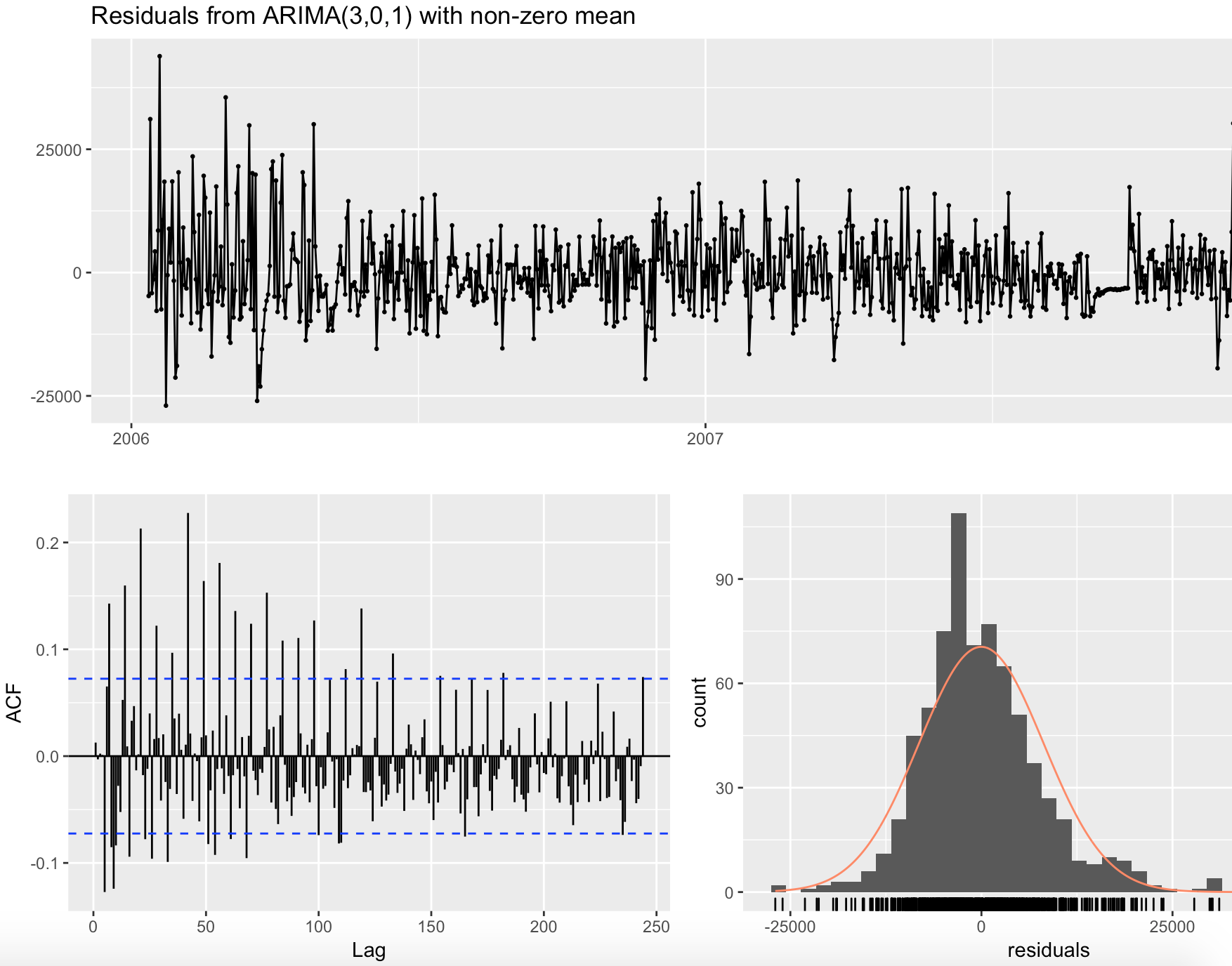


Daily Time Series Model - AIC

****

AIC value can estimate the quality of each model and relative to each of other models, lower AIC indicates a more parsimonious model, relative to a model fit with a higher AIC. Same as BIC indicator, which is a criterion for model selection among a finite set of models, the smaller BIC value the model is better. AIC value for daily time series model is too high, same as its BIC value. Therefore, daily time series model is not good to select.

Daily Time Series Model - Residual Diagnostics

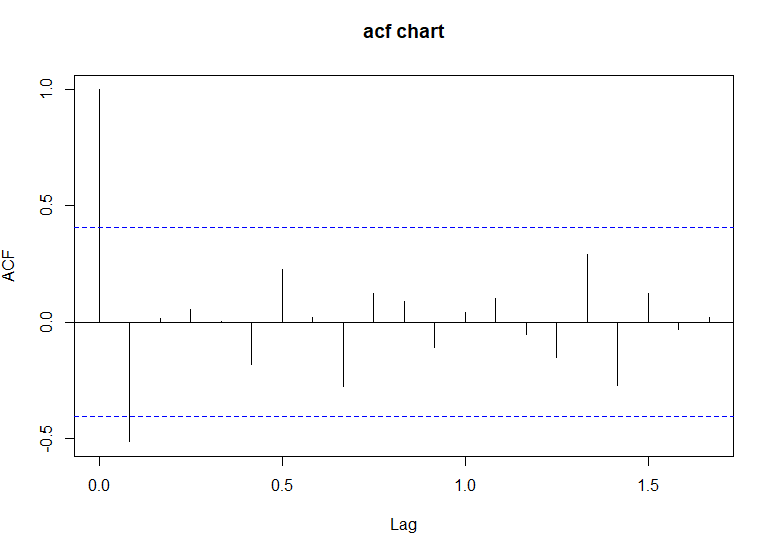
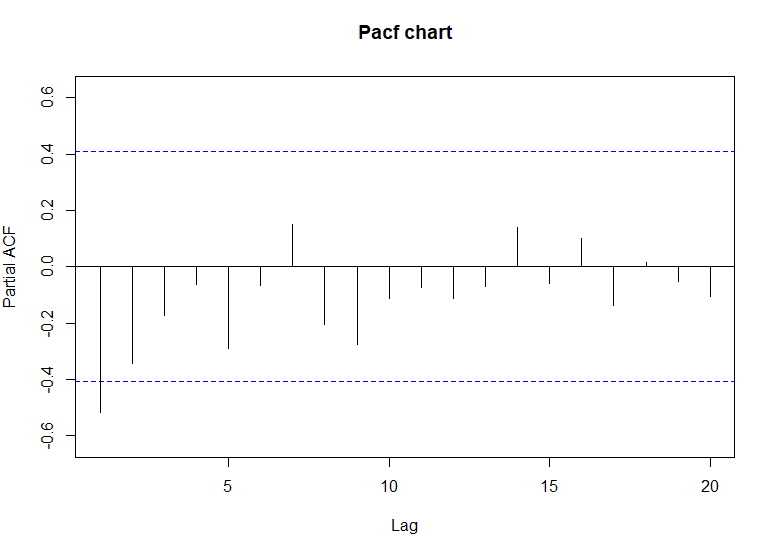


These graphs show the all residual information of daily time series model. The histogram of residuals shows its normal distribution, which is good but the ACF model gives the opposite results, there had significate correlation in the residuals series.

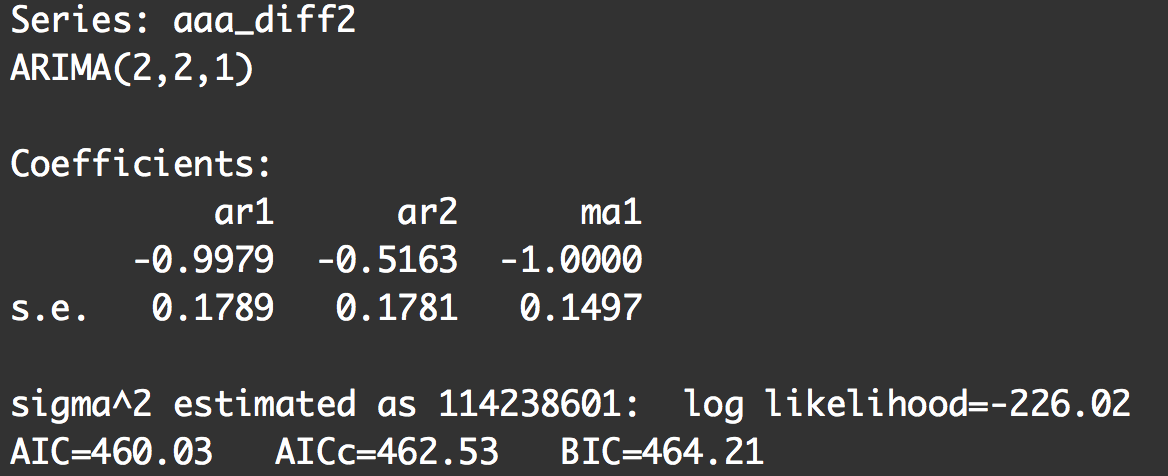
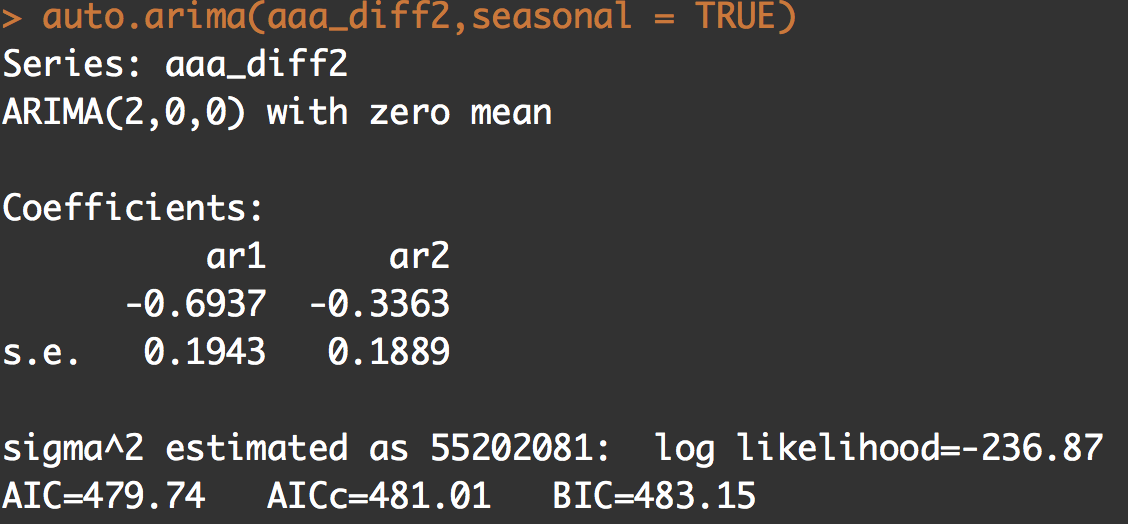
**Monthly Time Series Model Analysis**

Monthly Time Series Model - ACF and PACF

When ARIMA applied into monthly data, we select the model ARIMA(2,2,1). We also selected other similar model like ARIMA(2,2,0), which is generated by auto.arima function, and ARIMA(1,2,1). Only dose the model ARIMA(2,2,1) has the lowest AIC.

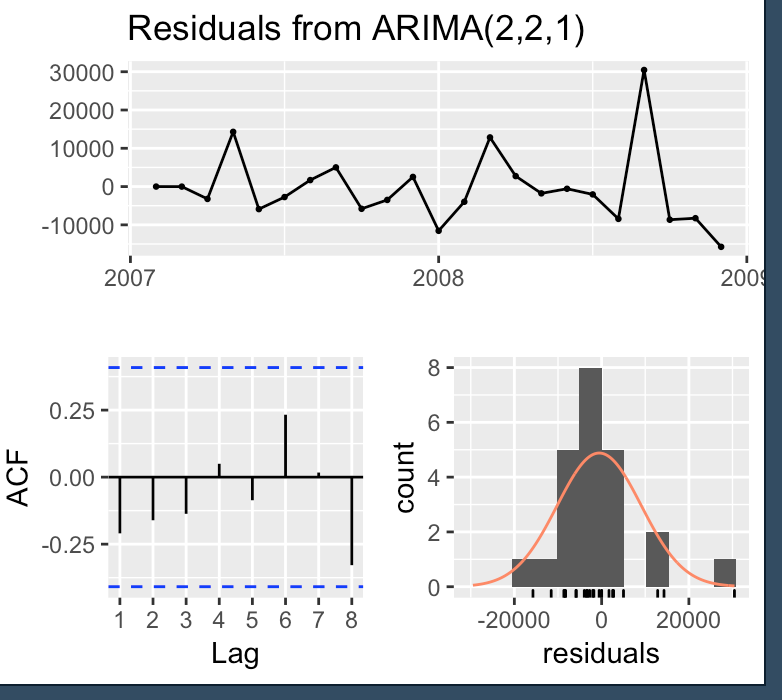


Monthly Time Series Model -AIC



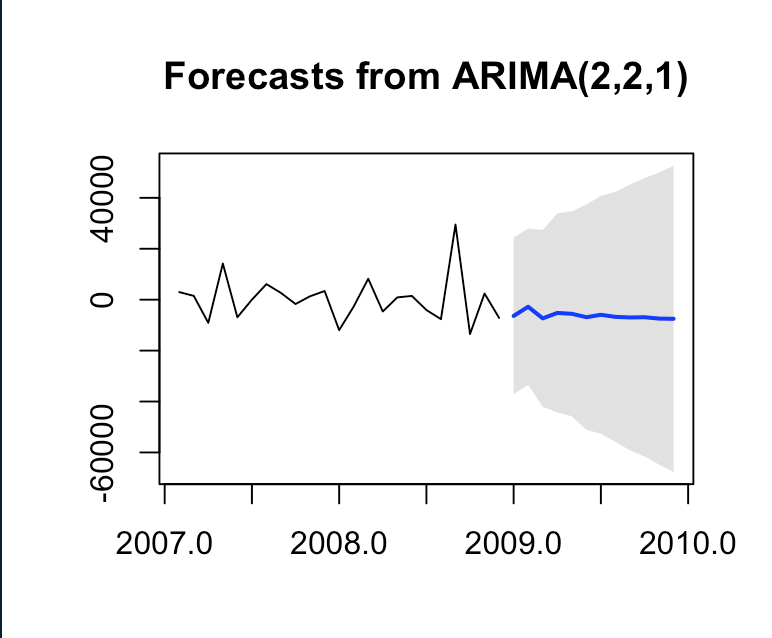
Compared model ARIMA (2,0,0) with model ARIMA (2,2,1), both AIC value and BIC of ARIMA (2,2,1) are lower than the former one. Therefore, it can be inferred that model ARIMA (2,2,1), which degree of differencing increased to 2 and the order of the moving-average model was improved to 1, is better than auto ARIMA model that has zero differencing degree and zero mean.

Monthly TS Model -Residual Diagnostics



These graphs show the all residual information of ARIMA (2,2,1) model. The mean of the residuals is close to zero means there is no significant correlation in the residuals series. What’s more, the time plot of the residuals shows that the variation of the residuals stays much the same across the historical data, apart from the one outlier, and therefore the residual variance can be treated as constant. This can also be seen on the histogram of the residuals. The histogram suggests that the residuals may be normal, forecasts from this method will probably be quite good.

**Predication for ARIMA (2,2,1)**



Based on the analysis above, model ARIMA (2,2,1) has the better performance than others, thus make the predication with this model. As the graph shows, usage of electricity reached to the peak in the third quarter of 2008 from year 2007 to 2009, and the straight line at the beginning of year 2009 shows an estimate electricity usage for the upcoming year. It is clear that the electricity usage will keep smooth and steady.

**Conclusion**

Both Daily T.S. model and monthly T.S. model can do time series analysis, while monthly T.S. model is better than daily T.S. model based on residual diagnostics analysis. AIC and BIC value help to choose the best model, which is used to do prediction. With analyzing AIC and BIC value, model ARIMA (2,2,1) has the better performance than the auto ARIMA.

Reference:

1. Time Series Analysis of Household Electric Consumption with ARIMA and ARMA Models. Retrieved from

http://www.iaeng.org/publication/IMECS2013/IMECS2013\_pp295-300.pdf

1. How to interpret acf and pacf plot. Retrieved from

https://stats.stackexchange.com/questions/147087/how-to-interpret-these-acf-and-pacf-plots

Appendix:

1. Code:

#loading necessary libraries

library(ggplot2)

library(dplyr)

library(forecast)

library(tseries)

library(TTR)

library(xts)

library(zoo)

#loading dataset

consumption <- read.csv(file.choose(),dec = ".")

#take a look

head(consumption)

str(consumption)

#Data cleansing

#There is missing value, consists of 1.25% of total dataset

#According to UCI desctiption, consecutive semicolons mean missing values.

#we will replace semicolon value by using previous value

#replace global\_active\_power

ms\_total <- consumption %>%

select(Date,Time,Global\_active\_power)

NAs <- ms\_total == "?"

[is.na(ms\_total)[NAs](is.na(ms_total)%5BNAs)] <- TRUE

ms\_total$Global\_active\_power <-

na.locf(ms\_total$Global\_active\_power, fromLast = FALSE)

consumption$Global\_active\_power <- ms\_total$Global\_active\_power

consumption %>%

select(Global\_active\_power) %>%

filter(Global\_active\_power == "?")

# na.loc is command to fill missing value.

#replace sub\_meter\_1

ms\_total <- consumption %>%

select(Date,Time,Sub\_metering\_1)

NAs <- ms\_total == "?"

[is.na(ms\_total)[NAs](is.na(ms_total)%5BNAs)] <- TRUE

ms\_total$Sub\_metering\_1 <-

na.locf(ms\_total$Sub\_metering\_1, fromLast = FALSE)

consumption$Sub\_metering\_1 <- ms\_total$Sub\_metering\_1

consumption %>%

select(Sub\_metering\_1) %>%

filter(Sub\_metering\_1 == "?")

#replace sub\_meter\_2

ms\_total <- consumption %>%

select(Date,Time,Sub\_metering\_2)

NAs <- ms\_total == "?"

[is.na(ms\_total)[NAs](is.na(ms_total)%5BNAs)] <- TRUE

ms\_total$Sub\_metering\_2 <-

na.locf(ms\_total$Sub\_metering\_2, fromLast = FALSE)

consumption$Sub\_metering\_2 <- ms\_total$Sub\_metering\_2

consumption %>%

select(Sub\_metering\_2) %>%

filter(Sub\_metering\_2 == "?")

#replace sum\_meter\_3

ms\_total <- consumption %>%

select(Date,Time,Sub\_metering\_3)

NAs <- ms\_total == "?"

[is.na(ms\_total)[NAs](is.na(ms_total)%5BNAs)] <- TRUE

ms\_total$Sub\_metering\_3 <-

na.locf(ms\_total$Sub\_metering\_3, fromLast = FALSE)

consumption$Sub\_metering\_3 <- ms\_total$Sub\_metering\_3

consumption %>%

select(Sub\_metering\_3) %>%

filter(Sub\_metering\_3 == "?")

#test semicolon value

consumption %>%

filter(Date == "28/4/2007")#We have deleted missing value

#Adding new columns and edit column names for clearance

#We will pay more attention on active energy consumption,

#Which means we will deleting Voltage,reactive power,intensity

#And we will add other appliance column for reference

consumption$Global\_active\_power <- as.numeric(levels(consumption$Global\_active\_power))[consumption$Global\_active\_power]

consumption$Sub\_metering\_1 <- as.numeric(levels(consumption$Sub\_metering\_1))[consumption$Sub\_metering\_1]

consumption$Sub\_metering\_2 <- as.numeric(levels(consumption$Sub\_metering\_2))[consumption$Sub\_metering\_2]

<is.na(consumption>)

head(consumption)

consumption$Date <- as.Date(consumption$Date,format = "%d/%m/%Y")

consumption$Time <- as.POSIXct(consumption$Time,format = "%H:%M:%S")

consumption <- consumption %>%

select(-Global\_reactive\_power,-Global\_intensity,-Voltage) %>%

mutate(total\_power = Global\_active\_power \* 1000/60)

consumption <- consumption %>%

mutate(other = total\_power -Sub\_metering\_1 -Sub\_metering\_2 - Sub\_metering\_3)

colnames(consumption)

consumption <- select(consumption, -Global\_active\_power)

colnames(consumption)[colnames(consumption) == "Sub\_metering\_1"] <- "Kitchen"

colnames(consumption)[colnames(consumption) == "Sub\_metering\_2"] <- "Laundry"

colnames(consumption)[colnames(consumption) == "Sub\_metering\_3"] <- "Living"

consumption <- select(consumption,Date,Time,total\_power,everything())

colnames(consumption)

#Draw 2006-12-16 all day electricity consumption

firstday <- consumption %>%

filter(Date == "2006-12-16")

str(firstday$Time)

ggplot(firstday,aes(x=Time,y=total\_power)) + geom\_line()

#get dataset grouped by date.

allyear <- consumption %>%

group\_by(Date) %>%

summarize\_at(.vars = vars(total\_power,Kitchen,Laundry,Living,other),

.funs = c(sum="sum"))

allyear$Living\_sum <- as.numeric(allyear$Living\_sum)

#draw plot based on allyear dataset

ggplot(allyear,aes(x=Date,y=total\_power\_sum)) + geom\_line() + stat\_smooth(color="red",fill="#FC4E07",method="loess")

#view of allyear

head(allyear)

tail(allyear)

allyear <- allyear[,c(1:6)]

str(allyear)

#convert into ts

#daily ts

#add three days to create complete two years period

new\_3days <- matrix(c(29322,2186,227,9271,17638,

29322,2186,227,9271,17638,

29322,2186,227,9271,17638),ncol = 5,byrow = TRUE)

allyear\_daily <- data.frame(date=seq(max(allyear$Date)+1, max(allyear$Date)+3,'days'), x = new\_3days )

tail(allyear\_daily)

colnames(allyear\_daily) <- c('Date','total\_power\_sum','Kitchen\_sum','Laundry\_sum','Living\_sum','other\_sum')

allyear\_daily <- rbind(allyear,allyear\_daily)

str(allyear\_daily)

daily\_ts <- ts(allyear\_daily$total\_power\_sum, frequency = 365.25, start=c(2006, 12,16))

plot(daily\_ts)

daily\_ts\_decompose <- decompose(daily\_ts)

plot(daily\_ts\_decompose)

adf.test(daily\_ts, alternative = 'stationary')

Pacf(daily\_ts,main = "Pacf chart",lag.max = 20)

acf(daily\_ts,main = "acf chart",lag.max = 50)

ddd <- unclass(daily\_ts)

#select model (3,0,1)

fit <- Arima(daily\_ts, order = c(3,0,1))

fit

checkresiduals(fit)

#monthly ts

agg\_Month <-aggregate(allyear,by=list(as.yearmon(allyear$Date,"%Y-%m-%D")), FUN=mean, na.rm=TRUE)

head(agg\_Month)

str(agg\_Month)

aaa <- ts(agg\_Month$total\_power\_sum, frequency = 12, start=c(2006, 12))

plot(aaa)

bbb <- decompose(aaa)

plot(bbb)

adf.test(aaa,alternative = "stationary")

aaa\_diff2 <-diff(aaa,differences=2)

adf.test(aaa\_diff2,alternative = "stationary")

Pacf(aaa\_diff2,main = "Pacf chart",lag.max = 20)

acf(aaa\_diff2,main = "acf chart",lag.max = 20)

auto.arima(aaa\_diff2,seasonal = TRUE)

#selece model (2,2,1)

fit2 <- Arima(aaa\_diff2, order = c(2,2,1))

fit2

checkresiduals(fit2)

#use monthly model to predict next one year

forecast\_fit2 <- forecast(fit2, h=12, level=c(99.5))

plot(forecast\_fit2)