EDA-681

Ran Tao, Ang Li, Yuan Yuan, Yukun He

2017/10/8

Introduction

In this project, our group studies the distribution of cooling degree days weekly total data of Florida during 2008, 2011 and 2014. We choose Florida as its higher demand for cooling throughout the year compared with other states in the US. Also, the data from Florida is mainly above zero, which is convenient for observation. Each year dataset includes 52 weeks of data. Comparing with monthly data, weekly data has larger amount of data which better represents the variation trend over a relatively long period of time.

Our goal is to estimate the distribution of the weekly total cooling degree days of Florida. By applying Explanatory Data Analysis, we estimate the distribution of Florida weekly total cooling degree days during 2008, 2011 and 2014. We start from the data reading, cleaning and manipulation by using loops, R packages such as dplyr and tidyverse. Then we visualize the data through histogram, boxplot, kernel density and ecdf. Besides, we use "modes" package to estimate the distribution, as the data follows a multimodal distribution. At last, we did comparison analysis of the distributions of 2008, 2011 and 2014.

Data Cleaning and Manipulation

```
library(ggplot2)
library(modes)

read_single_file<-function(file_name){
  #print(file_name)
  date = substr(file_name,nchar(file_name)-11,nchar(file_name)-4)
  text = readLines(con = file_name)
  data = text[skip_lines:length(text)]
  return (split_columns(data))
}

split_columns<-function(data){
  region_line = 52
  blank_line = 62
  regional_data = data.frame(matrix(nrow = length(data),ncol = 9))
  names(regional_data) = title
  for(i in 1:length(data)){
    new_line = unlist(strsplit(data[i],split = "(\\s\\s)+"))[1:10]
    data_line = data.frame(matrix(nrow = 1,ncol = 9))
    names(data_line) = title
    data_line[title][1] = new_line[1]
    data_line[title][2:length(data_line[title])] = as.numeric(new_line[2:length(new_line)])
    regional_data[i,] = data_line
  }
  #regional_data = regional_data[c(-region_line,-blank_line),]
  return(regional_data)
}
```

```
read_single_dir<-function(dir){
  file_list = list.files(path = dir,pattern = "*.txt")
  print(file_list)
  Florida = data.frame(matrix(nrow = 0,ncol = 9))
  names(Florida) = title
  for (file in file_list) {
    df = read_single_file(file_name = file.path(dir,file))
    rbind(Florida,df[10,])->Florida
    #assign(substr(file,1,nchar(file)-4),df,envir =.GlobalEnv)
  }
  row.names(Florida)<-substr(file_list,1,8)
  assign('Florida',Florida,envir = .GlobalEnv)
  return(Florida)
}

dir_list = list.dirs(recursive = FALSE)
title = c(
  "state","week total","week dev from norm","week dev from last year","cum total","cum dev from norm","
)
skip_lines = 16
Florida_year = list()
for(dir in dir_list){
  if(grepl(pattern = "*[12][0-9]{3}",x = dir)){
    F = read_single_dir(dir = dir)
    year = substr(dir,3,nchar(dir))
    Florida_year[[year]]<-F
  }
}

Florida$week=c(1:52)
Florida_2008<-as.data.frame(Florida_year[1])
Florida_2011<-as.data.frame(Florida_year[2])
Florida_2014<-as.data.frame(Florida_year[3])
```

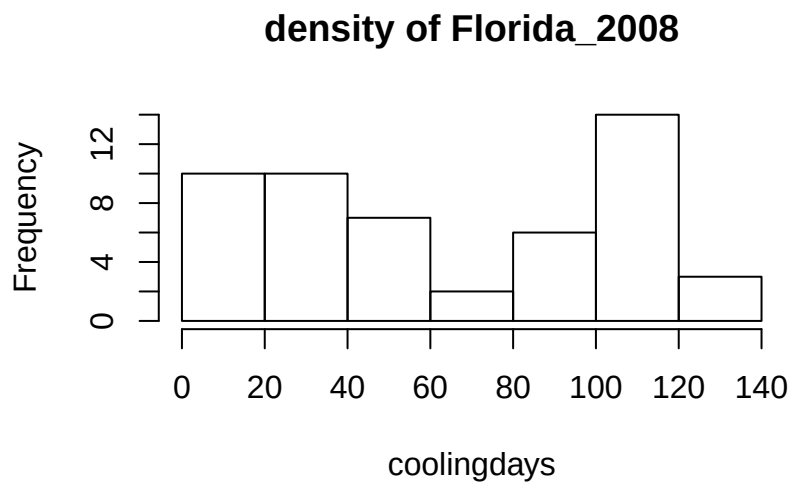
Data Visualization

1) 2008 Weekly Total Cooling Degree Days

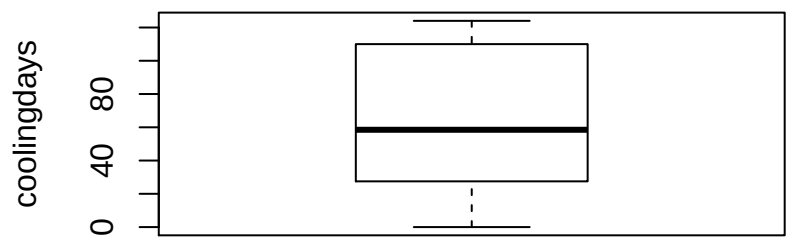
```
summary(Florida_2008$X2008.week.total)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   29.75   58.50   66.52  110.00  124.00
```

```
hist(Florida_2008$X2008.week.total,main = "density of Florida_2008",xlab = "coolingdays")
```



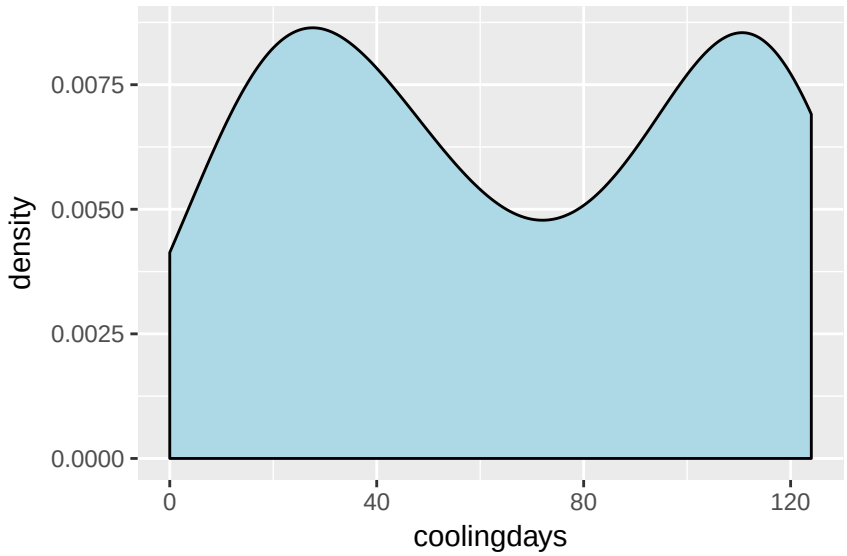

```
boxplot(Florida_2008$X2008.week.total,main = "boxplot of Florida_2008",ylab = "coolingdays")
```



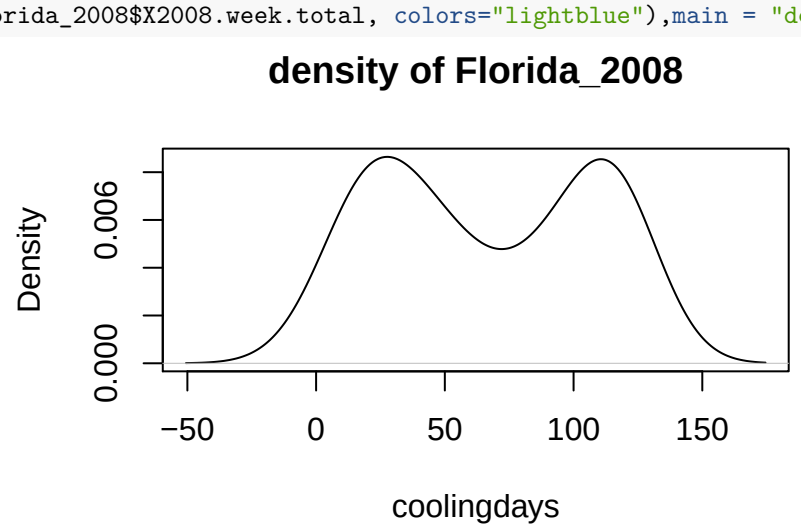

```
boxplot.stats(Florida_2008$X2008.week.total)
```

```
## $stats
## [1]   0.0  27.5  58.5 110.0 124.0
##
## $n
## [1] 52
##
## $conf
## [1] 40.42371 76.57629
##
## $out
## numeric(0)
```

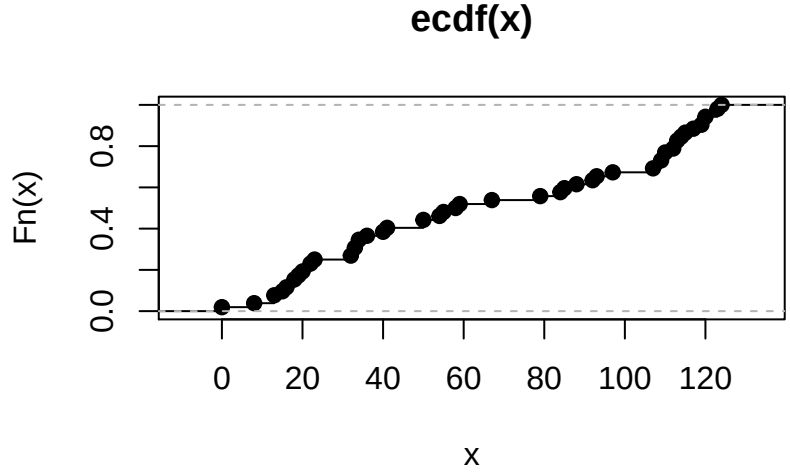
```
ggplot(data=Florida_2008, aes(x=Florida_2008$X2008.week.total))+geom_density(kernel="gaussian", fill="li
```




```
plot(density(Florida_2008$X2008.week.total, colors="lightblue"),main = "density of Florida_2008", xlab =
```




```
plot.ecdf(Florida_2008$X2008.week.total)
```



As we can see from the historgram and kernel density, the distribution of week total degree days in 2008 follows nonparametric distribution, as it has two peaks, it follows a multimodal distribution, which doesn't fit

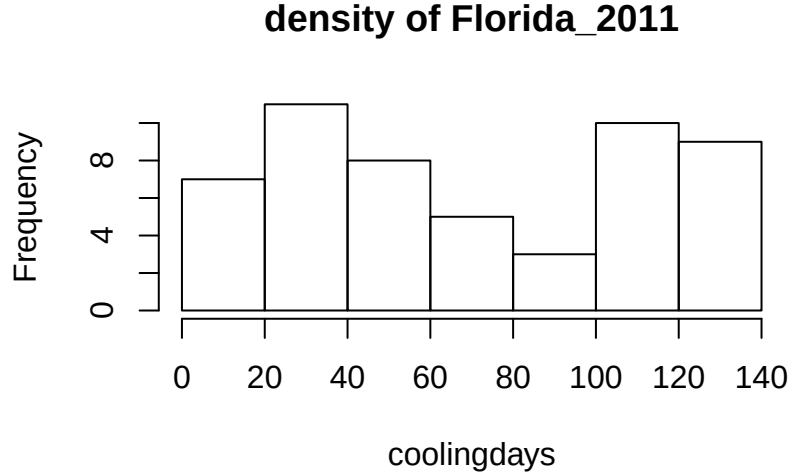
in any distribution model we have learned so far.

2) 2011 Weekly Total Cooling Degree Days

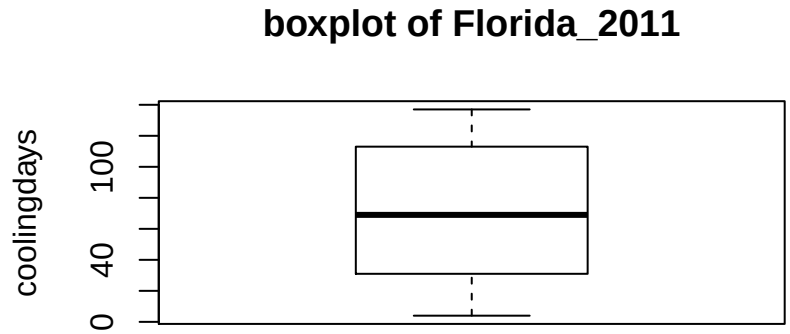
```
summary(Florida_2011$X2011.week.total)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     4.0    31.0    69.0    70.6   113.0   137.0
```

```
hist(Florida_2011$X2011.week.total,main = "density of Florida_2011",xlab = "coolingdays")
```



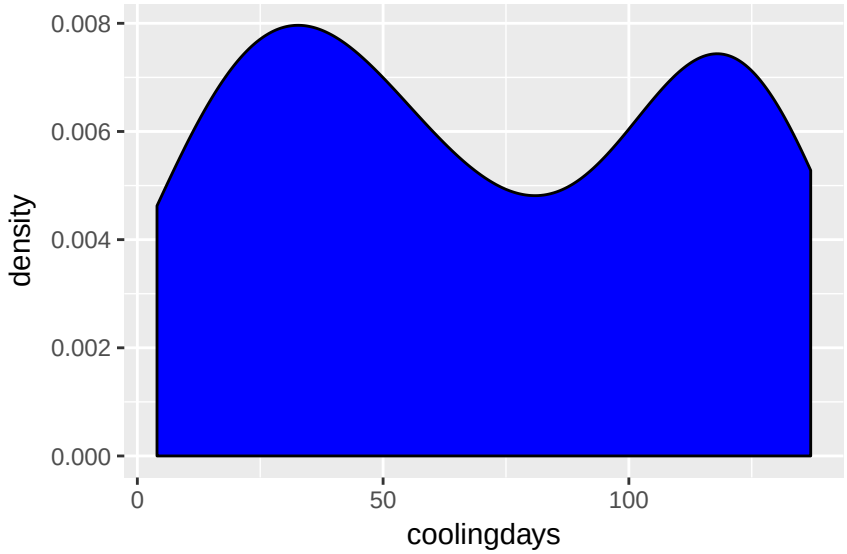

```
boxplot(Florida_2011$X2011.week.total,main = "boxplot of Florida_2011",ylab = "coolingdays")
```



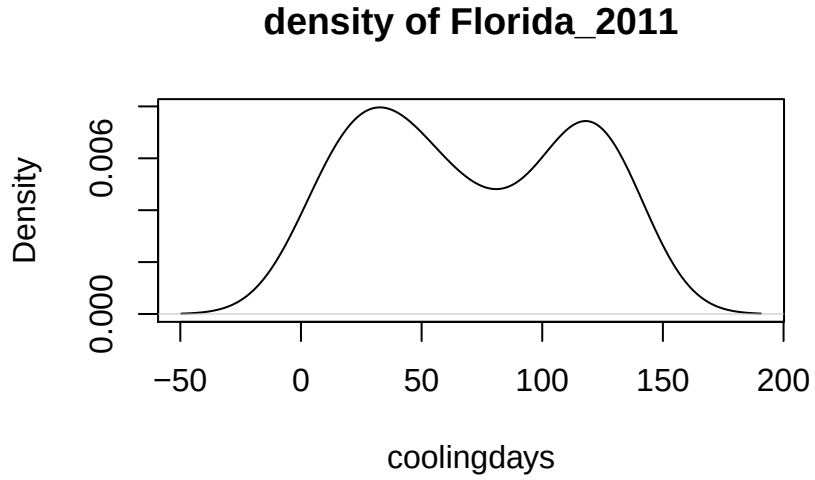

```
boxplot.stats(Florida_2011$X2011.week.total)
```

```
## $stats
## [1]   4  31  69 113 137
##
## $n
## [1] 53
##
## $conf
## [1] 51.20357 86.79643
##
## $out
## numeric(0)
```

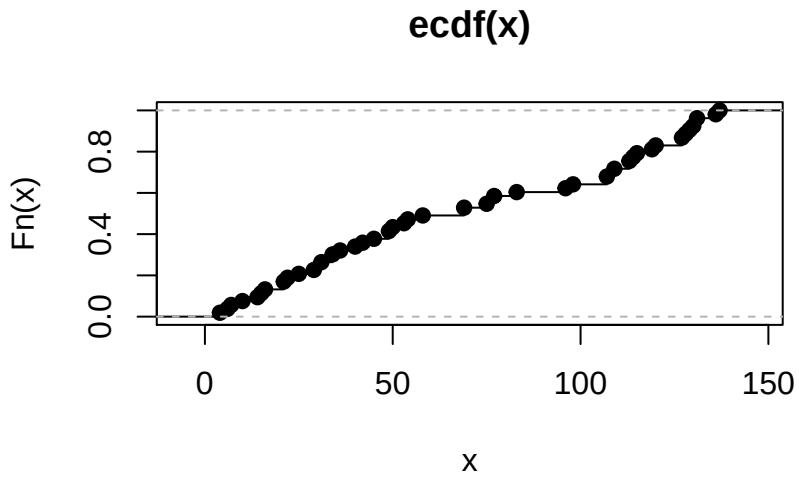
```
ggplot(data=Florida_2011, aes(x=Florida_2011$X2011.week.total))+geom_density(kernel="gaussian", fill="bl
```




```
plot(density(Florida_2011$X2011.week.total),main = "density of Florida_2011", xlab = "coolingdays")
```




```
plot.ecdf(Florida_2011$X2011.week.total)
```



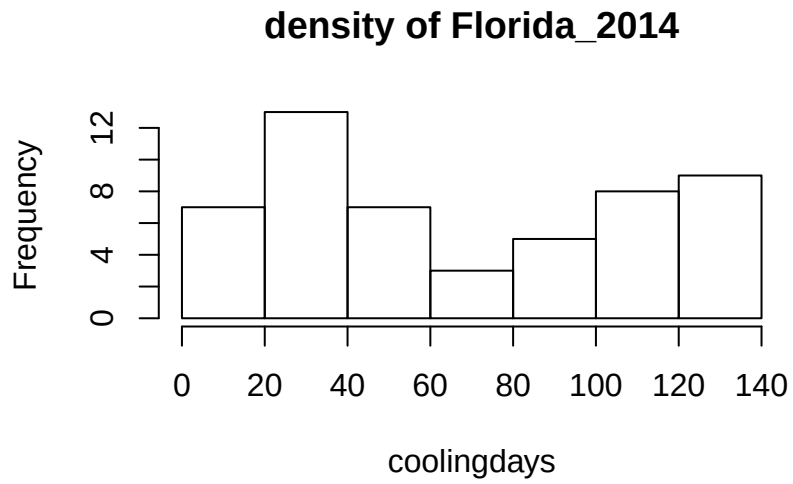
As we can see from the historgram and kernel density, the distribution of week total degree days in 2011 follows nonparametric distribution, as it has two peaks, it follows a multimodal distribution, which is similar to that of 2008

3) 2014 Weekly Total Cooling Degree Days

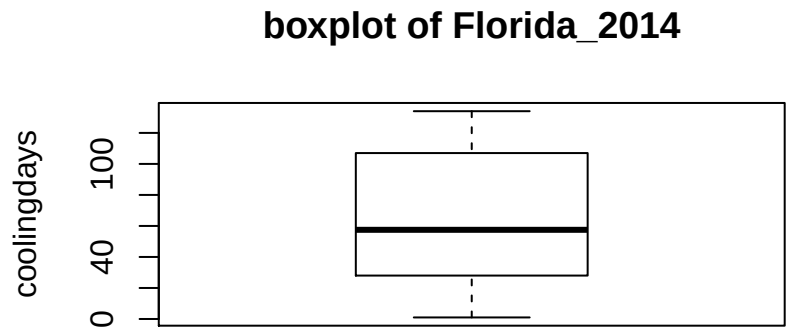
```
summary(Florida_2014$X2014.week.total)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   28.50   57.50   67.15  106.00  134.00
```

```
hist(Florida_2014$X2014.week.total,main = "density of Florida_2014",xlab = "coolingdays")
```



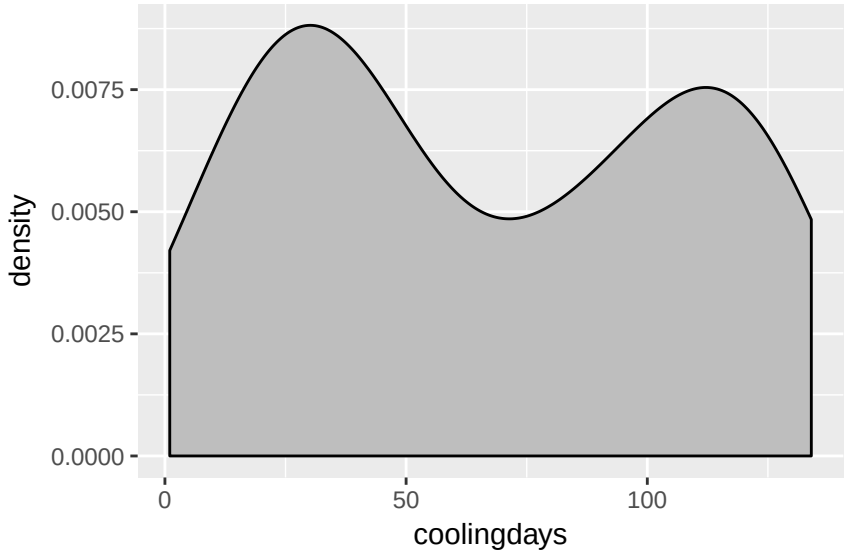

```
boxplot(Florida_2014$X2014.week.total, main = "boxplot of Florida_2014", ylab = "coolingdays")
```



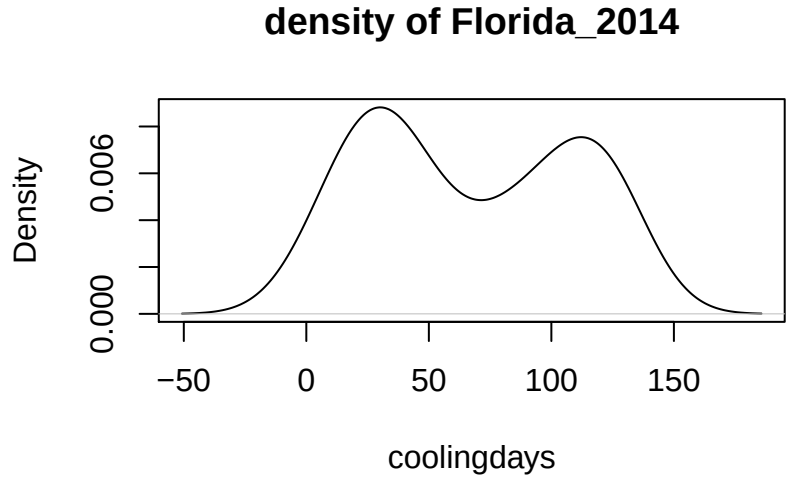

```
boxplot.stats(Florida_2014$X2014.week.total)
```

```
## $stats
## [1]   1.0  28.0  57.5 107.0 134.0
##
## $n
## [1] 52
##
## $conf
## [1] 40.19058 74.80942
##
## $out
## numeric(0)
```

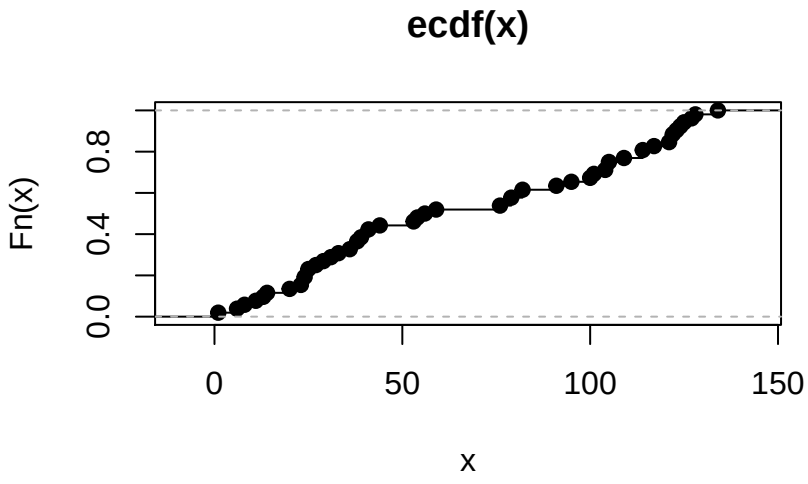
```
ggplot(data=Florida_2014, aes(x=Florida_2014$X2014.week.total))+geom_density(kernel="gaussian", fill="g
```




```
plot(density(Florida_2014$X2014.week.total),main = "density of Florida_2014", xlab = "coolingdays")
```



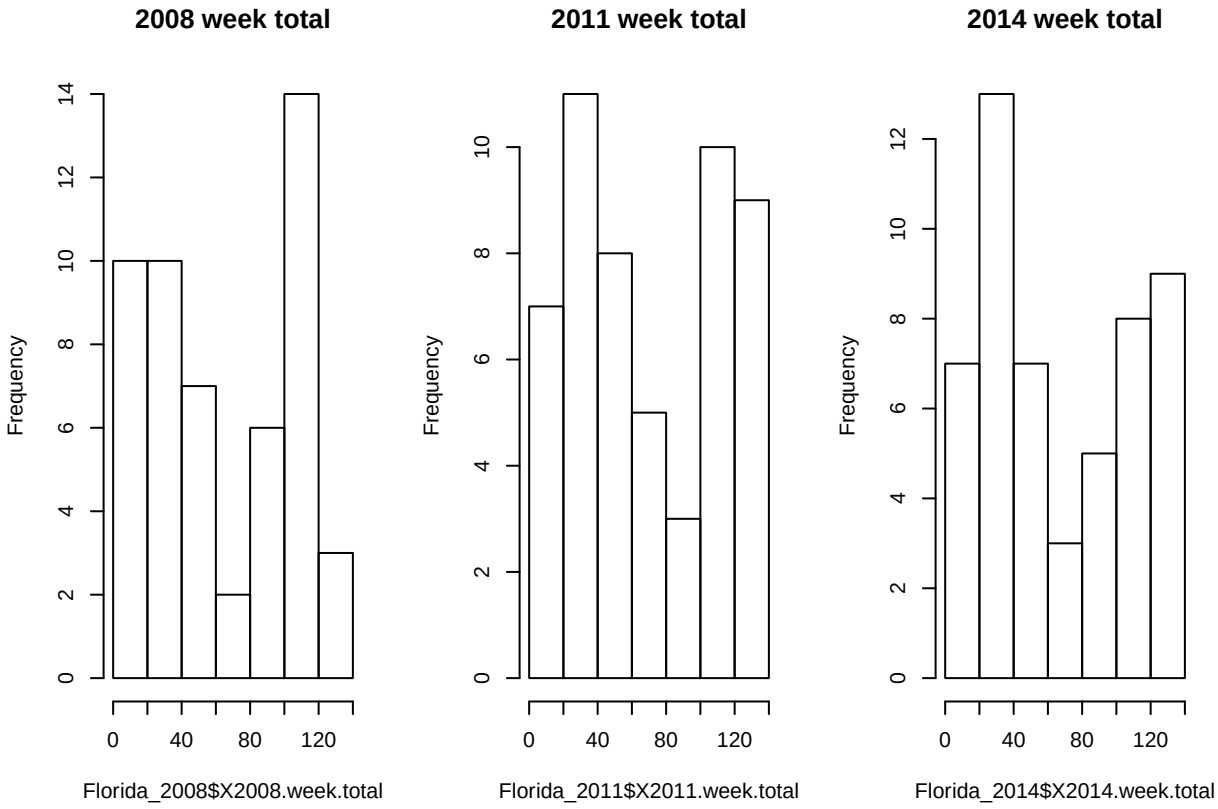

```
plot.ecdf(Florida_2014$X2014.week.total)
```



As we can see from the historgram and kernel density, the distribution of week total degree days in 2014 follows nonparametric distribution, as it has two peaks, it follows a multimodal distribution, which is similar to those of 2008 and 2011

Histgram of Weekly Total Degree Day Comparison For 3 years

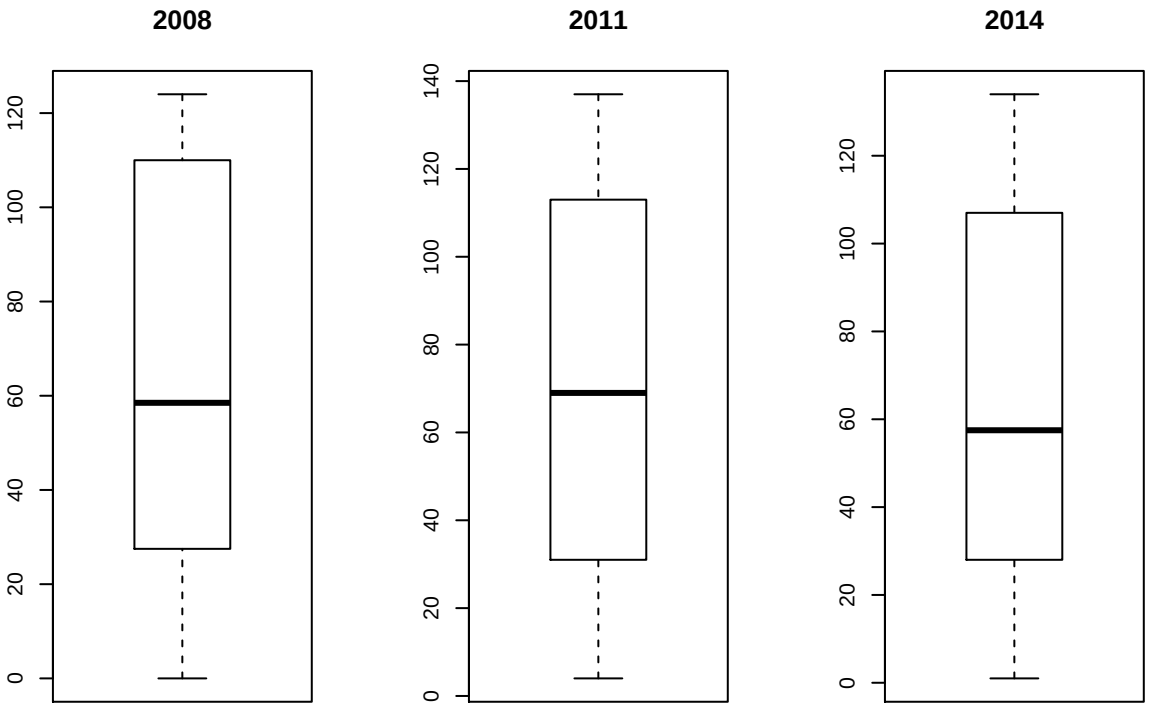
```
par(mfrow=c(1,3))
hist(Florida_2008$X2008.week.total, main="2008 week total")
hist(Florida_2011$X2011.week.total, main="2011 week total")
hist(Florida_2014$X2014.week.total, main="2014 week total")
```



In this comparison of histagram, we notice that the density of week total degree days for 3 years are similar. And since 2011, there were more degree days that is higher than 120.

Boxplot of Weekly Total Degree Day Comparison For 3 years

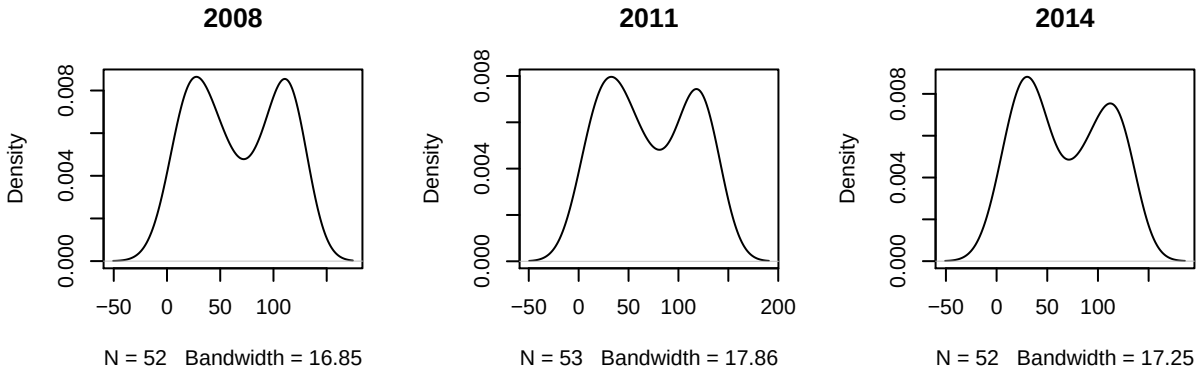
```
par(mfrow=c(1,3))
boxplot(Florida_2008$X2008.week.total,main= "2008")
boxplot(Florida_2011$X2011.week.total, main="2011")
boxplot(Florida_2014$X2014.week.total, main="2014")
```



We noticed that 2011 has a little higher mean for weekly total degree days, compare to other two years.

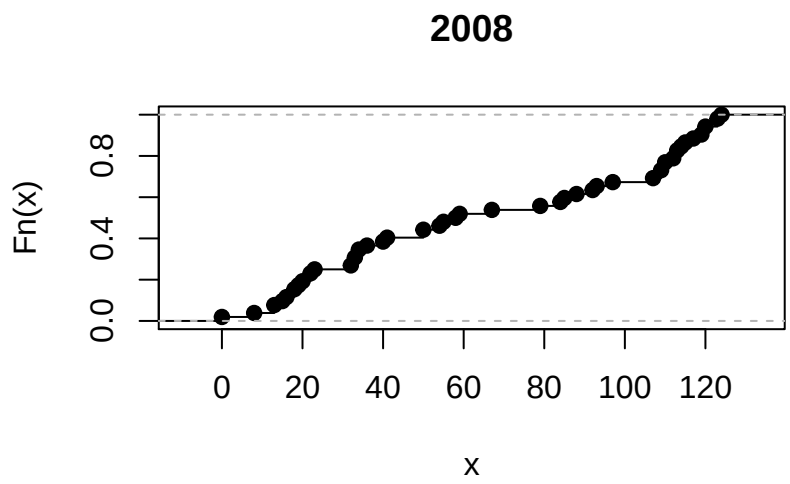
Density Plot of Weekly Total Degree Day Comparison For 3 years

```
par(mfrow=c(2,3))
plot(density(Florida_2008$X2008.week.total), main = "2008")
plot(density(Florida_2011$X2011.week.total), main = "2011")
plot(density(Florida_2014$X2014.week.total), main = "2014")
```

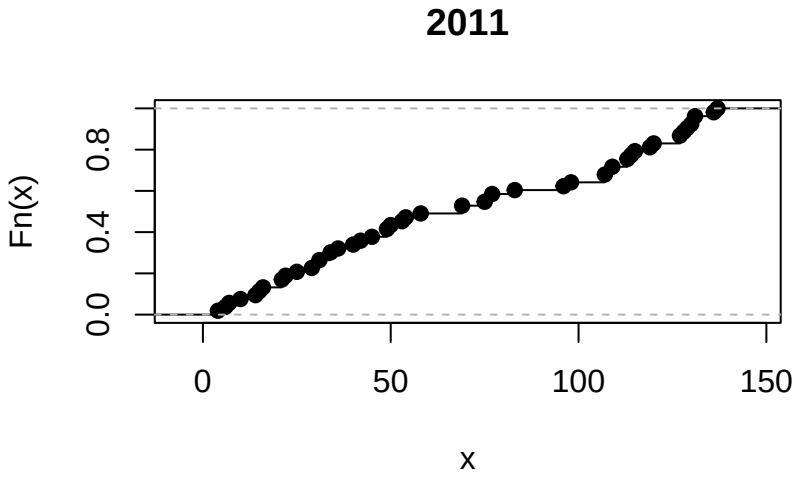


ECDF Plot of weekly Total Degree Day Comparison For 3 years

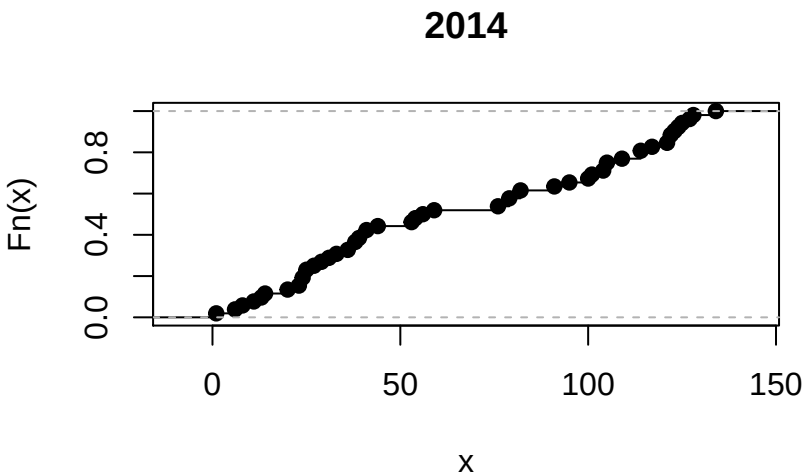
```
plot.ecdf(Florida_2008$X2008.week.total, main="2008")
```




```
plot.ecdf(Florida_2011$X2011.week.total, main="2011")
```




```
plot.ecdf(Florida_2014$X2014.week.total, main="2014")
```



As shown in density plot, this is a multimodal distribution,

Which leads to our next move—we use “modes” package in r to try to estimate the distribution of week total degree days.

Model Estimation

```
find_modes<-function(x){
  modes<-NULL
  for (i in 2:(length(x)-1)){
  if(x[i]>x[i-1] & (x[i]>x[i+1])){
    modes<-c(modes,i)
  }
  }
  if(length(modes)==0){
    modes='This is monotonic distribution'
  }
  return(modes)
}

mymodes_indices2008<-find_modes(density(Florida_2008$X2008.week.total)$y)
mymodes_indices2011<-find_modes(density(Florida_2011$X2011.week.total)$y)
mymodes_indices2014<-find_modes(density(Florida_2014$X2014.week.total)$y)

Var2008<-density(Florida_2008$X2008.week.total)$y[mymodes_indices2008]
Peak2008<-density(Florida_2008$X2008.week.total)$x[mymodes_indices2008]

Var2011<-density(Florida_2011$X2011.week.total)$y[mymodes_indices2011]
Peak2011<-density(Florida_2011$X2011.week.total)$x[mymodes_indices2011]

Var2014<-density(Florida_2014$X2014.week.total)$y[mymodes_indices2014]
Peak2014<-density(Florida_2014$X2014.week.total)$x[mymodes_indices2014]

Var2008
```

```
## [1] 0.008641451 0.008542786
```

```
Peak2008
```

```
## [1]  27.85901 110.67845
```

```
Var2011
```

```
## [1] 0.007962986 0.007436022
```

```
Peak2011
```

```
## [1]  32.66185 118.20897
```

```
Var2014
```

```
## [1] 0.008815867 0.007544155
```

```
Peak2014
```

```
## [1]  30.24083 112.16472
```

For year 2008, 2011, and 2014, what we observe from the diagrams are two peak values each year, respectively locate at 27.86 and 110.68 unit, 32.66 and 118.21 unit, and 30.24 and 112.16 unit. These represent the levels of needs of cooling during per year. The diagrams also introduce a novel data distribution - bimodal distribution - to our sights. Moreover, both data of the lower peak and higher peak are increasing year by year, which represents a trend of increasing temperature year by year in Florida. We think this is a phenomenon of Global Warming.

Furthermore, we also record the corresponding variances of six modes, which are 0.86% and 0.85% for 2008, 0.79% and 0.74% for 2011, and 0.88% and 0.75% for 2014. These relatively small variances reveal the fact that more data fall in the range around the peak value, which represent an overall high demand of cooling throughout the year on the other hand.

Conclusion

After all, it not hard to conclude that all three years datasets separately follow bimodal distributions, which is still novel and mysterious for us to comprehensively interpret. Bimodal distribution is a very common data distribution in the field of nature science. However, it is rarely studied because of the difficulties in estimating its parameters either with frequentist or Bayesian methods. One thing to be sure of is that the mean value of the distribution can be a more robust sample estimator than the median of the distribution. And we can also assume the data are from a combination of two normal distributions.

So far, our observation and analysis reveal the fact that although data varies year by year, data reach distinct local maximum twice each year. We can conclude that Florida is a state with high temperature on average, and the demand for cooling concentrate on two levels throughout a year.