615 Midterm Project

Tianwen Huan

10/21/2016

```
CT <- read.csv("~/Desktop/GlobalLandTemperaturesByCity.csv")
```

1. Check the original dataset

```
summary(CT)
```

```
##           dt          AverageTemperature AverageTemperatureUncertainty
##  1882-01-01:   3510   Min.   :-42.7      Min.   : 0.0
##  1882-02-01:   3510   1st Qu.: 10.3      1st Qu.: 0.3
##  1882-03-01:   3510   Median : 18.8      Median : 0.6
##  1882-04-01:   3510   Mean   : 16.7      Mean   : 1.0
##  1882-05-01:   3510   3rd Qu.: 25.2      3rd Qu.: 1.3
##  1882-06-01:   3510   Max.   : 39.7      Max.   :15.4
##  (Other)   :8578152   NA's   :364130     NA's   :364130
##           City                 Country           Latitude
##  Springfield:   9545   India         :1014906   36.17N : 425455
##  Worcester  :   8359   China         : 827802   34.56N : 351472
##  León       :   7469   United States: 687289   52.24N : 347775
##  Rongcheng  :   6526   Brazil        : 475580   40.99N : 331559
##  Birmingham :   6478   Russia        : 461234   23.31N : 319266
##  Brest      :   6478   Japan         : 358669   50.63N : 308886
##  (Other)    :8554357   (Other)       :4773732   (Other):6514799
##    Longitude
##  139.23E: 129600
##  88.25E :  88842
##  136.22E:  86940
##  0.00W  :  83557
##  46.31W :  82878
##  5.26E  :  64780
##  (Other):8062615
```

```
dim(CT)
```

```
## [1] 8599212       7
```

```
str(CT$Country)
```

```
##  Factor w/ 159 levels "Afghanistan",..: 40 40 40 40 40 40 40 40 40 40 ...
```

Due to the huge original dataset(time form 1743.11 to 2013.09, 8599211 rows, 159 Countries), we decide to choose the data from 1900.01 only for the United States, which also includes 350805 observations.

2. Choose the subset

```
cityT <- CT %>%
     filter(Country=="United States") %>% # narrow down to United States
     mutate(date=dt) %>%
     separate(dt, c("year", "month", "day")) %>% # sepearate the year month and day
     filter(year >= 1900)  # select the data after year 1990

# drop the "day" and "Country" columns
cityT <- subset(cityT, select = c(10,1,2,4,5,6,8,9))

# check missing data
cityT[!complete.cases(cityT),]
```

```
##             date year month AverageTemperature
## 10920 2013-09-01 2013    09                 NA
##       AverageTemperatureUncertainty      City Latitude Longitude
## 10920                            NA Anchorage   61.88N   151.13W
```

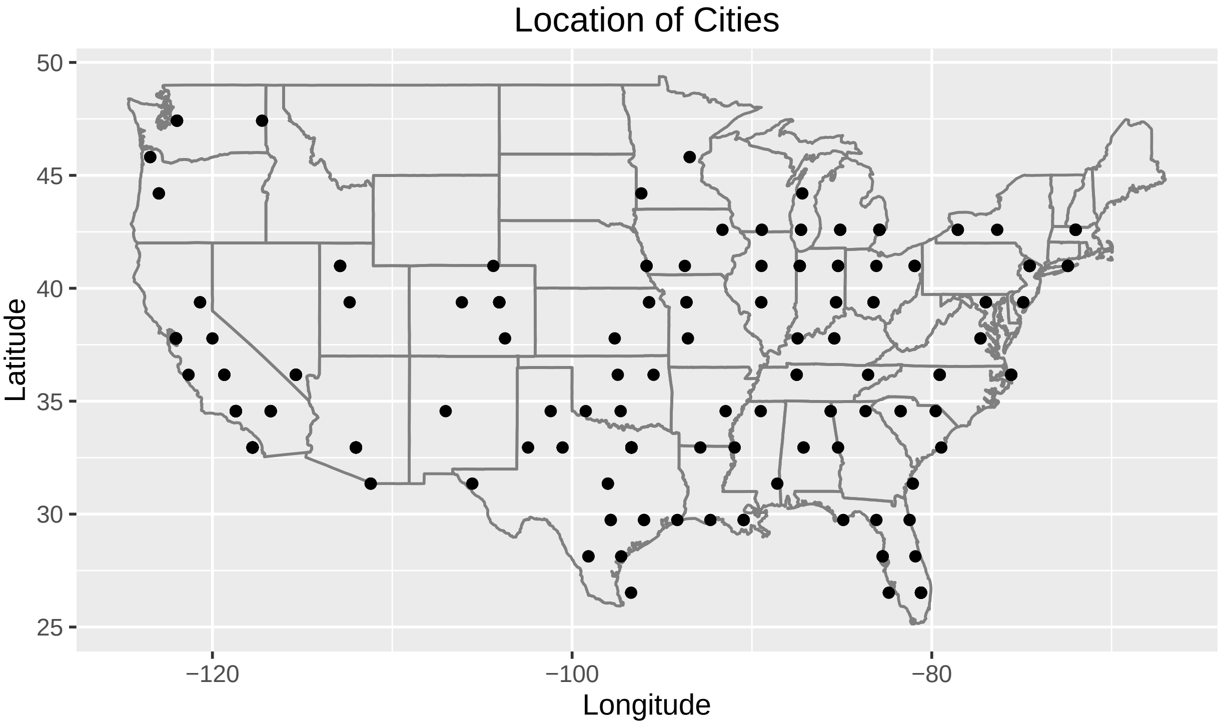
```
# drop all the data for 2013.09
cityT <- cityT %>%
        filter(year!="2013" | month!="09")

write.csv(cityT, 'cityT.csv')
```

3. Data character transformation

```
cityT$date<-as.Date(cityT$date,"%Y-%m-%d")
cityT$year<-as.numeric(cityT$year,"%Y")
cityT$month<-as.numeric(cityT$month,"%m")


cityT$lat<-as.numeric(gsub("N|E|S|W", "",cityT$Latitude))*ifelse(grepl("S",cityT$Latitude),-1,1)
cityT$long<-as.numeric(gsub("N|E|S|W", "", cityT$Longitude))*ifelse(grepl("W",cityT$Longitude),-1,1)

cityT <- data.table(cityT)

# remove Hawaii & Alaska
cityT <- cityT %>%
        filter(long>=-130 & lat>=25 & lat<=55)
```

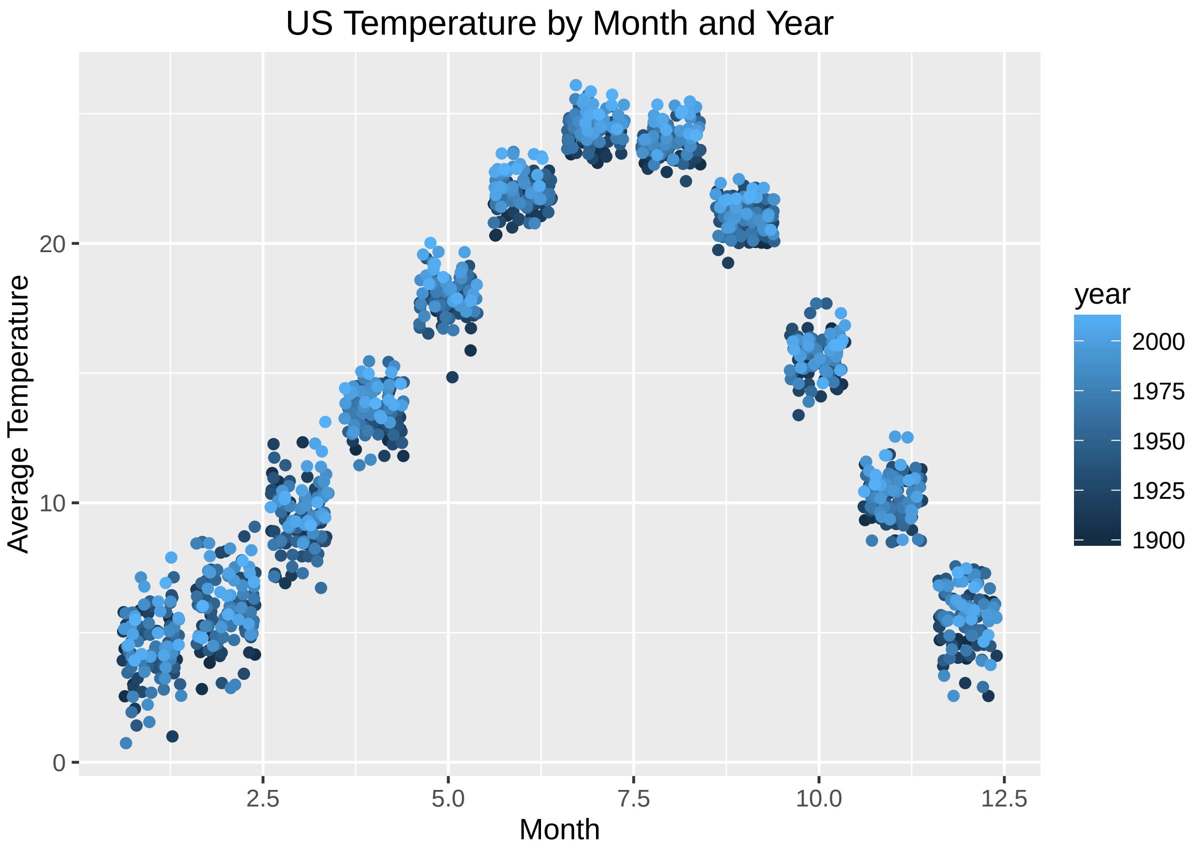
4. Location Graph According to latitudes and longitude

```
citylocation <- subset(cityT, select = c("City","lat","long"))
citylocation <- citylocation %>% distinct(.keep_all= FALSE)


ggplot(citylocation, aes(long, lat), col=temp) +
  borders("state") + geom_point()+
  scale_size_area() + coord_quickmap() +
  labs(x="Longitude", y="Latitude", title="Location of Cities")
```

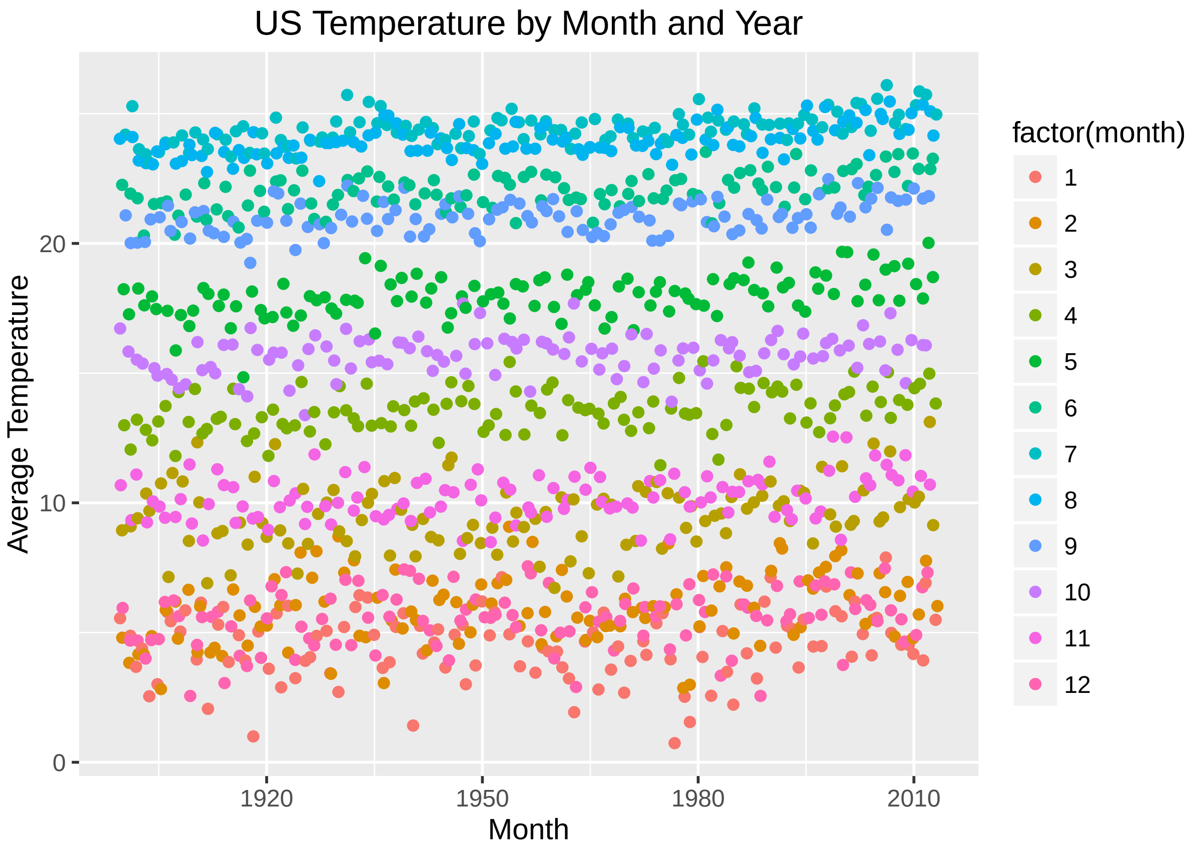


5. US Temperature by Month and Year 1900-2012

```r
# choose subset
aT <- cityT %>%
       group_by(year, month) %>%
       summarise(temp=mean(AverageTemperature))

# month trend for different year
ggplot(aT, aes(x=month, y=temp)) +
  geom_jitter(aes(colour=year)) + ggtitle("US Temperature by Month and Year") +
  labs(x="Month", y="Average Temperature")
```

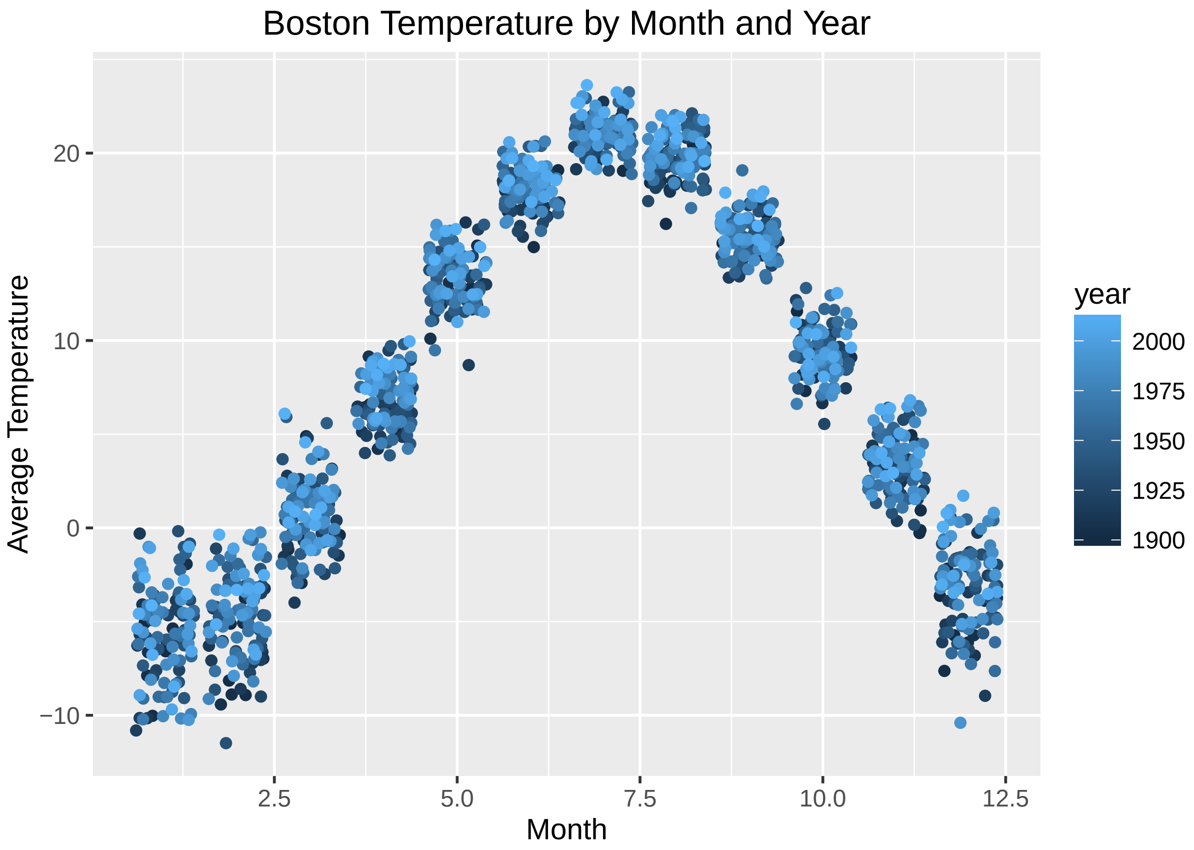


```r
# year trend for different month
ggplot(aT, aes(x=year, y=temp)) +
  geom_jitter(aes(colour=factor(month))) +
  ggtitle("US Temperature by Month and Year") +
  labs(x="Month", y="Average Temperature")
```

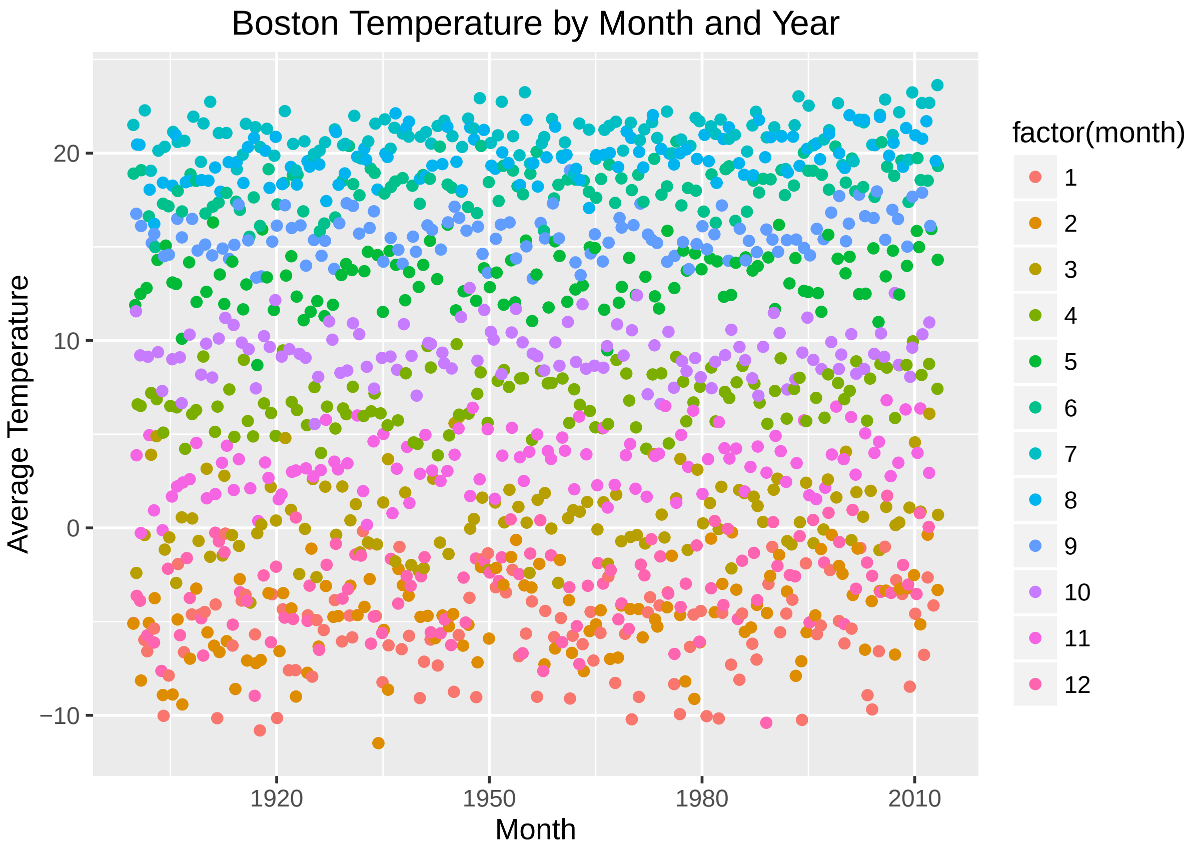


6. Boston Temperature by Month and Year 1900-2012

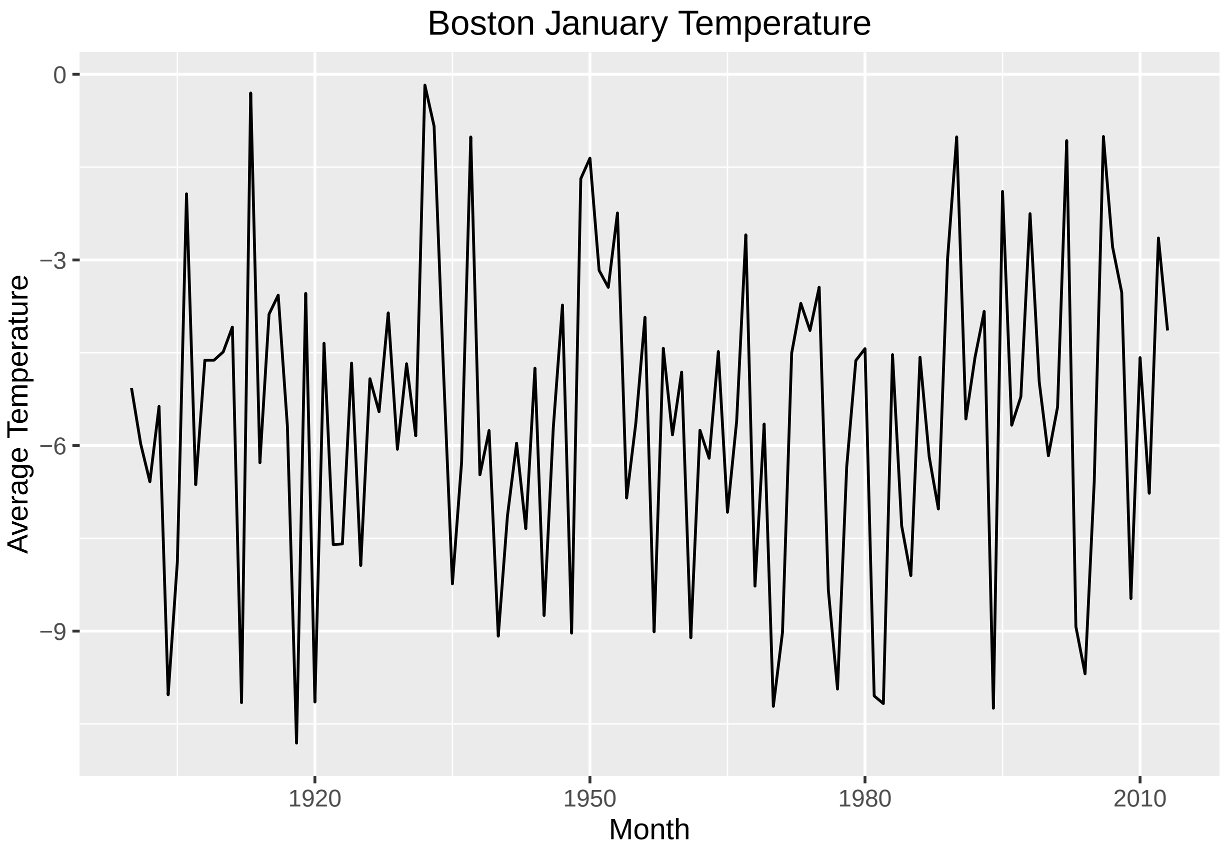
```
# choose subset
aTB <- cityT %>%
       filter(City=="Boston") %>%
       group_by(year, month)

# month trend for different year
ggplot(aTB, aes(x=month, y=AverageTemperature)) +
  geom_jitter(aes(colour=year)) +
  ggtitle("Boston Temperature by Month and Year") +
  labs(x="Month", y="Average Temperature")
```



```
# year trend for different month
ggplot(aTB, aes(x=year, y=AverageTemperature)) +
  geom_jitter(aes(colour=factor(month))) +
  ggtitle("Boston Temperature by Month and Year") +
  labs(x="Month", y="Average Temperature")
```

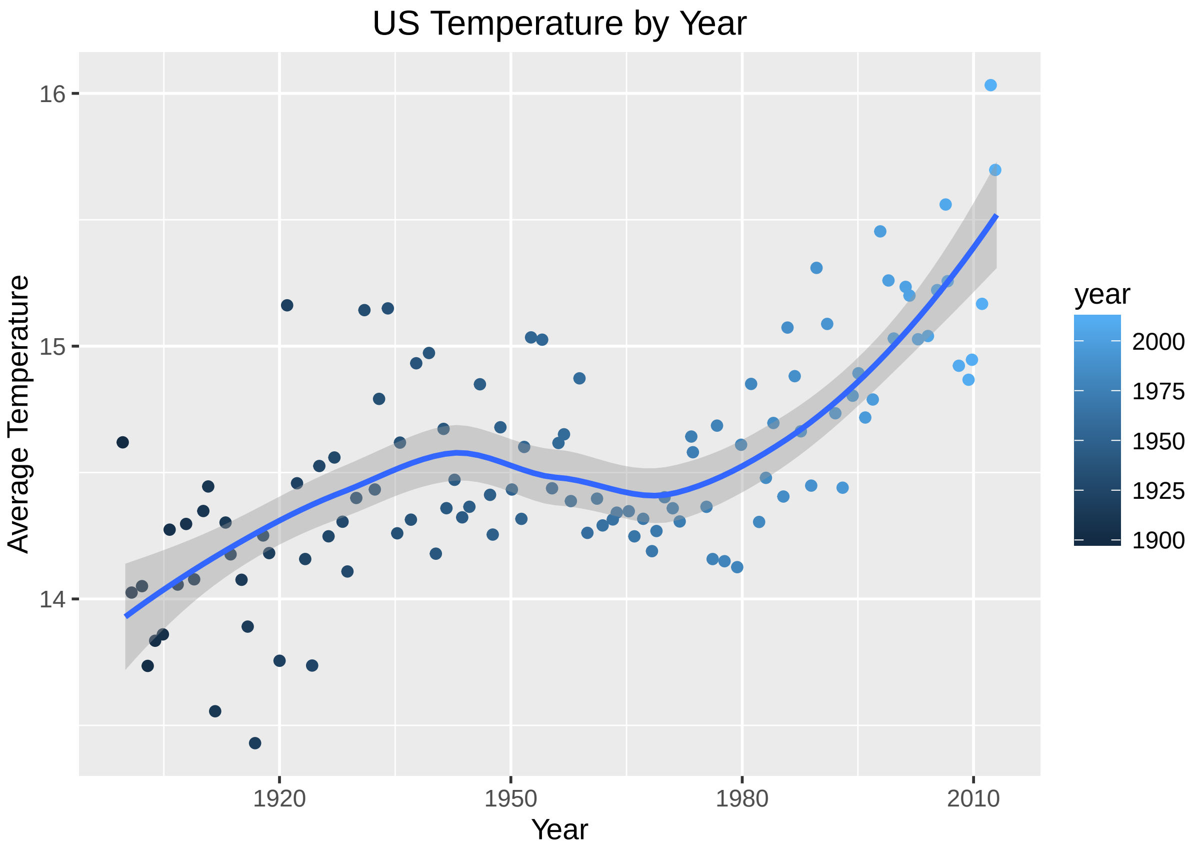


```
# year trend for January
BostonT<- cityT %>% filter(City == "Boston")
bostonjan<-BostonT %>% filter(month == 1)
ggplot(bostonjan, aes(x=year, y=AverageTemperature))+
  geom_line() +
  labs(x="Month", y="Average Temperature", title="Boston January Temperature")
```



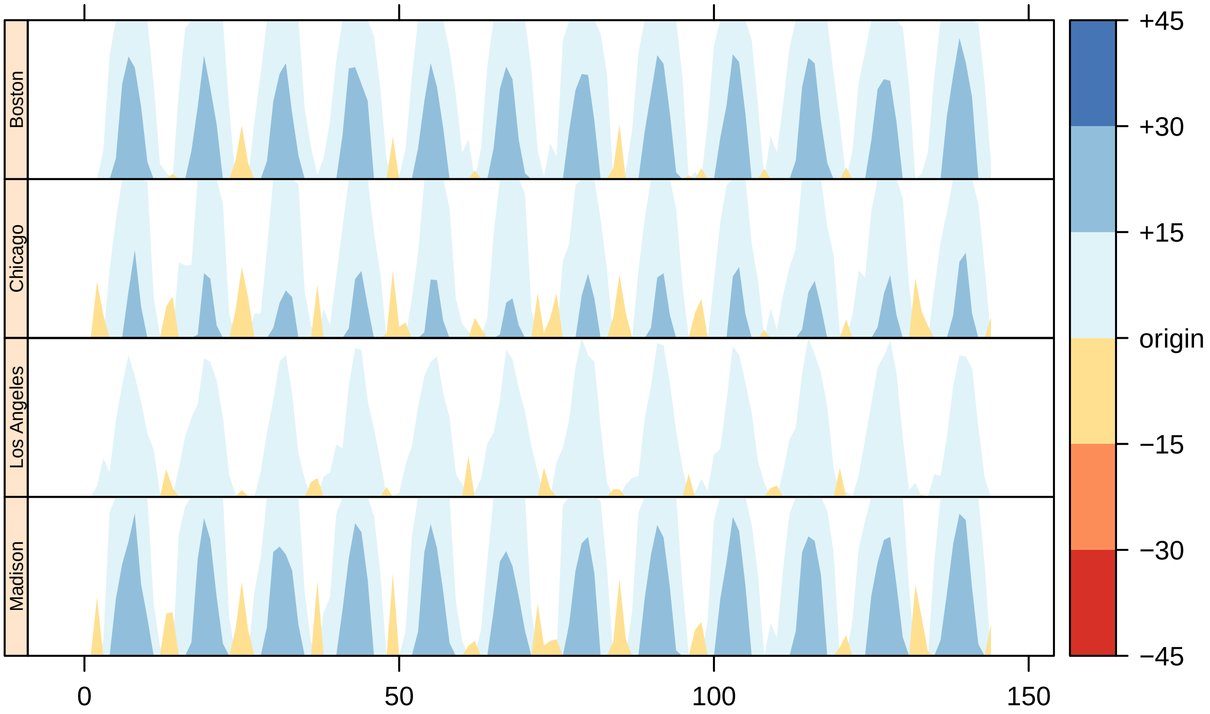
7. US Temperature by Year 1900-2012

```
# choose subset
aTy <- cityT %>%
       group_by(year) %>%
       summarise(temp=mean(AverageTemperature))

# year trend
ggplot(aTy, aes(x=year, y=temp)) +
  geom_jitter(aes(colour=year)) +
  ggtitle("US Temperature by Year") + geom_smooth() +
  labs(x="Year", y="Average Temperature")
```



8. Horizonplot

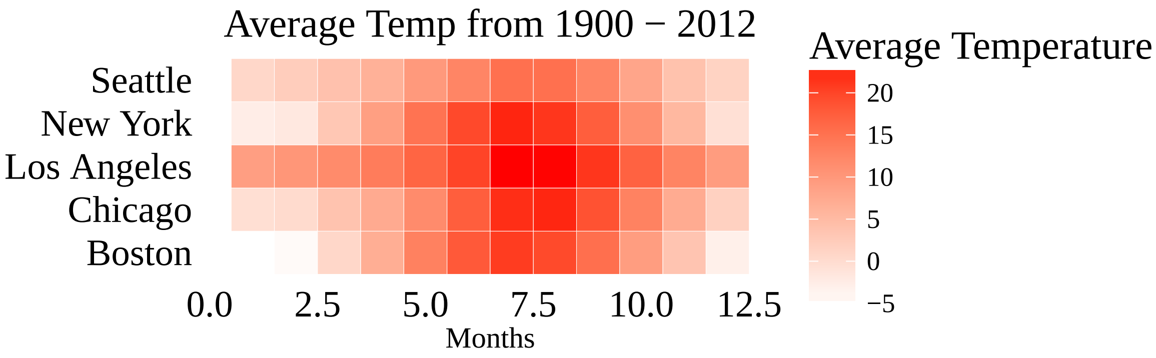
```
y10 <- subset(cityT, year %in% c(1900,1910,1920,1930,1940,1950,1960,1970,1980,1990,2000,2010))

citytable1 <- y10 %>%
             filter(City=="Boston" | City=="Chicago" | City=="Los Angeles" | City=="Madison") %>%
             group_by(year, month, City) %>%
             summarise(temp=AverageTemperature) %>%
             spread(City, temp)

citytable <- subset(citytable1, select=c(3:6))

horizonplot(ts(citytable), horizonscale = 15, colorkey = TRUE,
            xlab="1900,1910,1920,1930,1940,1950,1960,1970,1980,1990,2000,2010")
```



1900,1910,1920,1930,1940,1950,1960,1970,1980,1990,2000,2010

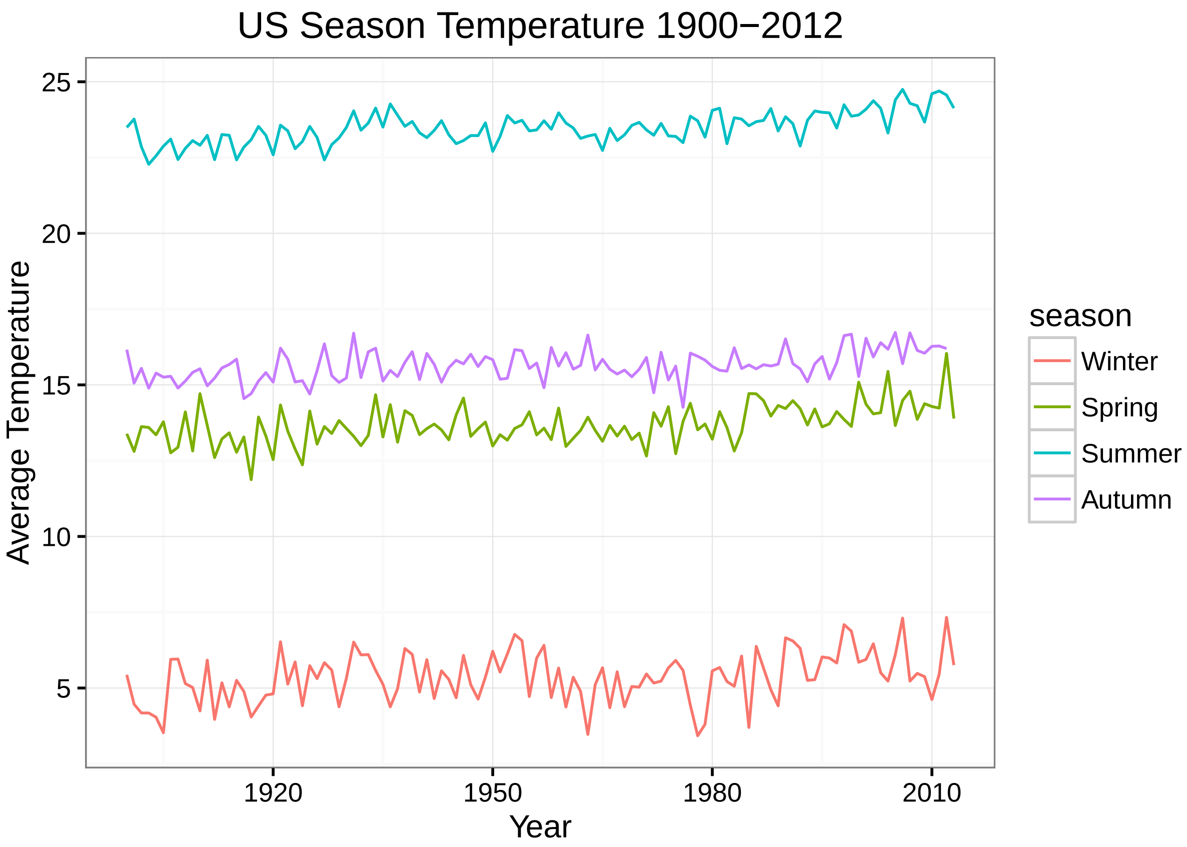
9. Heatmap for 5 main Cities

```
# choose the subset
hm <- cityT %>%
        group_by(month, City) %>%
        summarise(temp=mean(AverageTemperature)) %>%
        filter(City=="Boston" | City=="Chicago" |
                 City=="Los Angeles" | City=="New York" | City=="Seattle")

# Heatmap
ggplot(hm, aes(x=month, y=City, fill=temp, frame=City)) +
  geom_tile(color="white", size=0.1) +
  scale_fill_gradient(name="Average Temperature", low="white", high="red") +
  coord_equal() +
  labs(x = "Months", y = "", title = "Average Temp from 1900 - 2012") +
  theme_tufte() +
  theme(axis.ticks = element_blank()) +
  theme(axis.text = element_text(size = 14)) +
  theme(plot.title = element_text(size = 15)) +
  theme(legend.title = element_text(size = 15)) +
  theme(legend.text = element_text(size = 10))
```

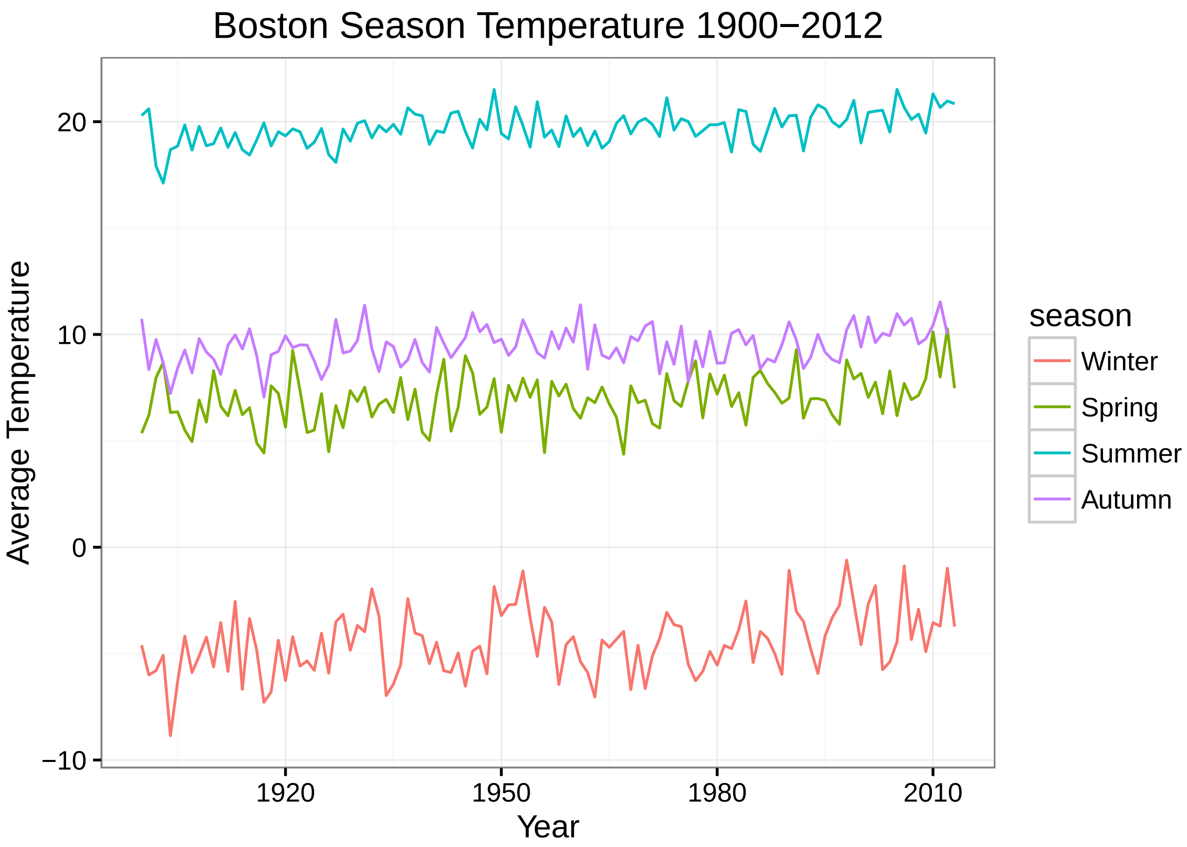


10. Temperature Trend by Season

```
s <- function(m){
            factor((m %% 12) %/% 3, labels = c('Winter', 'Spring', 'Summer', 'Autumn'))}

par(mfrow=c(1,2))
# US Tempreture by Season
sus <- cityT %>%
       group_by(year, season=s(month)) %>%
       summarise(temp=mean(AverageTemperature))

sus %>%
    ggplot(mapping=aes(x=year, y=temp)) +
     geom_line(mapping = aes(color=season)) +
     theme_bw() + labs(x="Year", y="Average Temperature",
                       title="US Season Temperature 1900-2012")
```



```
# Boston Tempreture by Season
sbo <- cityT %>%
       filter(City=="Boston") %>%
       group_by(year, season=s(month)) %>%
       summarise(temp=mean(AverageTemperature))

sbo %>%
    ggplot(mapping=aes(x=year, y=temp)) +
     geom_line(mapping = aes(color=season)) +
     theme_bw() + labs(x="Year", y="Average Temperature",
                       title="Boston Season Temperature 1900-2012")
```


Boston Season Temperature 1900−2012
Average Temperature
20
10
0
−10
1920
1950
1980
2010
Year
season
Winter
Spring
Summer
Autumn