Climate Analysis in R

Caio di Felice Cunha

Analyzing the temperature in R

For this project, we will analyze the database provided by Berkeley Earth. In version 1, we will restrict ourselves to the United States area and see only a few cities, with succinct conclusions

Stage 1 - Import the libraries and connect with SQL

Here is the data collection, in this case a csv file downloaded from http://berkeleyearth.org/data I uploaded this file and insert into MySQL Database Server

```
## Loading the necessary libraries:
  library(viridis)
  library(hrbrthemes)
  library(plotly)
  library(readr)
  library(dplyr)
  library(ggplot2)
  library(scales)
  library(data.table)
  library(tibble)
  library(RMySQL)

## Creating connection with MySQL
con = dbConnect(
  MySQL(),
  user = "root",
  password = a,
  dbname = db,
  host = ht)
```

Stage 2 - Selecting the data properly

For this particular project, I just looked into North American temperatures

Therefore the Country column is no longer necessary, because it would be all ‘United States’. Also Latitude and Longitude.

```
qry <- "
  select
    dt,
    AverageTemperature,
    AverageTemperatureUncertainty,
    City
  from
    globalclimate
  where
    Country = 'United States';
"

US_Climate <- dbGetQuery(con, qry)
```

Stage 3 - Preparating and Organizating Data

```
## Converting Data and creating Month and Year column
US_Climate$dt <- as.POSIXct(US_Climate$dt, format = "%Y-%m-%d")
US_Climate$Month <- month(US_Climate$dt)
US_Climate$Year <- year(US_Climate$dt)

## Percentage of null values
sum(is.na(US_Climate)) / nrow(US_Climate)

## [1] 0.08395013

## We have 8% of the data as "na", so I chose to just delete them
US_Climate <- na.omit(US_Climate)
sum(is.na(US_Climate)) / nrow(US_Climate)

## [1] 0
```

When looking at the overall temperature graph, we notice that it assumes what we call the “Left Skewed Distribution”, as temperatures tend to be greater than 0

Stage 4 - Chosing the States Capital

```
## Selecting the capital of the states
Capital_Cities <-
  US_Climate[US_Climate$City %in%
               c('Montgomery','Juneau','Phoenix','Little Rock',
'Sacramento',
                 'Denver','Hartford', 'Dover','Tallahassee','Atlanta',
                 'Honolulu', 'Boise','Springfield', 'Indianapolis',
                 'Des Moines', 'Topeka', 'Frankfort', 'Baton Rouge',
'Augusta',
                 'Annapolis', 'Boston','Lansing','Saint Paul', 'Jackson',
                 'Jefferson City', 'Helena','Lincoln','Carson
City','Concord',
                 'Trenton','Santa
Fe','Albany','Raleigh','Bismarck','Columbus',
                 'Oklahoma
City','Salem','Harrisburg','Providence','Columbia',
                 'Pierre','Nashville','Austin','Salt Lake
City','Montpelier',

'Richmond','Olympia','Charleston','Madison','Cheyenne'),]
```

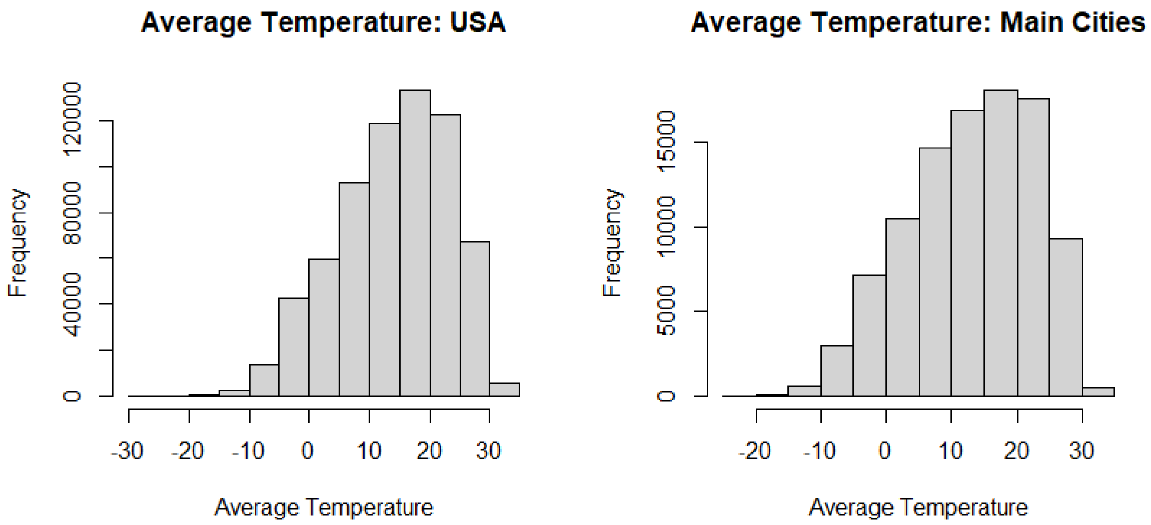
Stage 5 - Analyze the two datasets

```
## Analyze the two datasets
summary(US_Climate)

##        dt                         AverageTemperature
##  Min.   :1743-11-01 00:00:00.00   Min.   :-25.16    
##  1st Qu.:1852-09-01 00:00:00.00   1st Qu.:  7.82    
##  Median :1906-02-01 00:00:00.00   Median : 14.96    
##  Mean   :1902-03-09 22:07:39.44   Mean   : 13.97    
##  3rd Qu.:1960-01-01 00:00:00.00   3rd Qu.: 21.10    
##  Max.   :2013-09-01 00:00:00.00   Max.   : 34.38    
##  AverageTemperatureUncertainty     City               Month       
##  Min.   : 0.040                Length:659468      Min.   : 1.000  
##  1st Qu.: 0.300                Class  :character  1st Qu.: 3.000  
##  Median : 0.530                Mode   :character  Median : 6.000  
##  Mean   : 1.092                                   Mean   : 6.481  
##  3rd Qu.: 1.650                                   3rd Qu.: 9.000  
##  Max.   :10.520                                   Max.   :12.000  
##       Year     
##  Min.   :1743  
##  1st Qu.:1852  
##  Median :1906  
##  Mean   :1902  
##  3rd Qu.:1960  
##  Max.   :2013  

summary(Capital_Cities)

##        dt                         AverageTemperature
##  Min.   :1743-11-01 00:00:00.00   Min.   :-21.99    
##  1st Qu.:1836-03-01 00:00:00.00   1st Qu.:  6.25    
##  Median :1895-12-01 00:00:00.00   Median : 13.88    
##  Mean   :1893-02-21 02:48:30.32   Mean   : 13.02    
##  3rd Qu.:1954-12-01 00:00:00.00   3rd Qu.: 20.73    
##  Max.   :2013-09-01 00:00:00.00   Max.   : 34.38    
##  AverageTemperatureUncertainty     City               Month      
##  Min.   : 0.040                Length:98410       Min.   : 1.00  
##  1st Qu.: 0.300                Class  :character  1st Qu.: 3.00  
##  Median : 0.550                Mode   :character  Median : 6.00  
##  Mean   : 1.221                                   Mean   : 6.48  
##  3rd Qu.: 1.980                                   3rd Qu.: 9.00  
##  Max.   :10.520                                   Max.   :12.00  
##       Year     
##  Min.   :1743  
##  1st Qu.:1836  
##  Median :1895  
##  Mean   :1893  
##  3rd Qu.:1954  
##  Max.   :2013  
```

As you can see, there is almost no difference between the datasets, so we are going to work with the "Capital_Cities" in order to better look to data

Another way to confirm that is to look to a histogram

```
par(mfrow=c(1,2))
hist(
  US_Climate$AverageTemperature,
  main = "Average Temperature: USA",
  xlab = "Average Temperature"
)

hist(
  Capital_Cities$AverageTemperature,
  main = "Average Temperature: Main Cities",
  xlab = "Average Temperature"
)
```


Average Temperature: USA

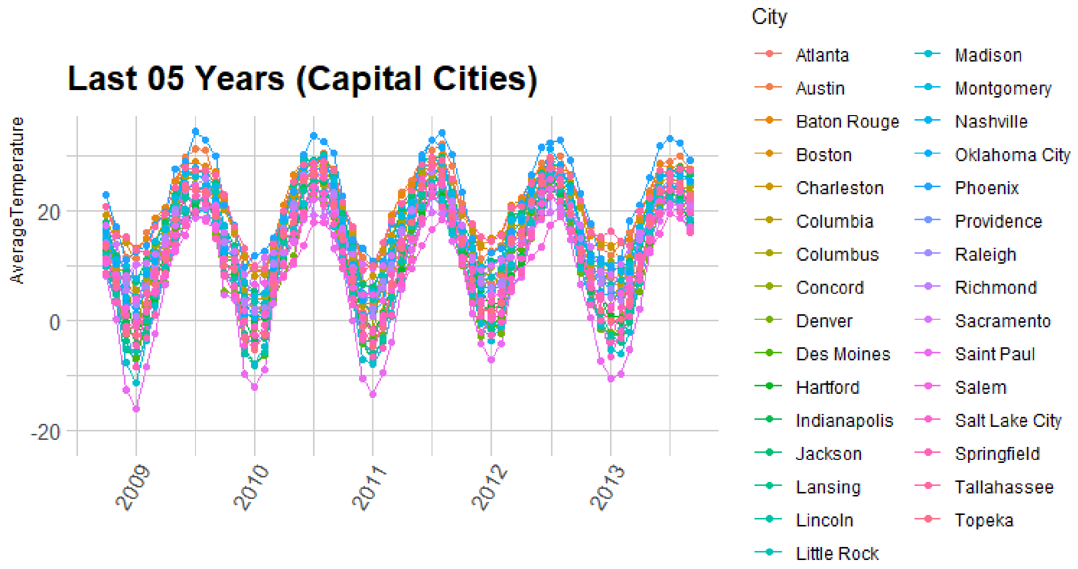
Average Temperature: Main Cities

```
par(mfrow=c(1,1))
```

The data distribution is similar

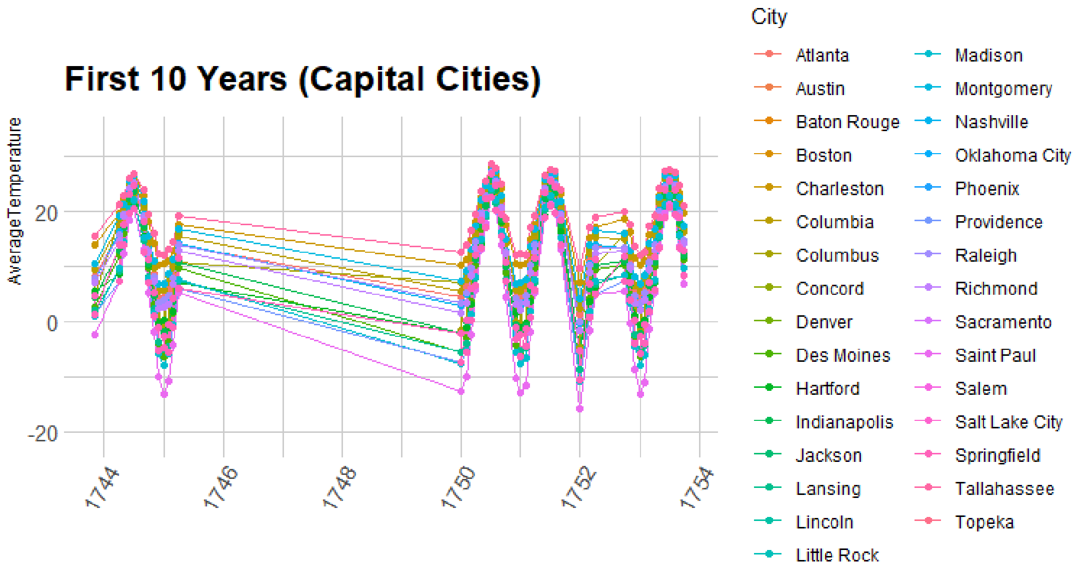
Stage 6 - Analysis over time

```
Capital_Cities$dt <- as.Date(Capital_Cities$dt)

ggplot(Capital_Cities,
            aes(x=dt,
                y=AverageTemperature,
                color = City,
                group = City)) +
  geom_line() +
  geom_point() +
  xlab("") +
  theme_ipsum() +
  theme(
    axis.text.x = element_text(angle=60, hjust=1)) +
  scale_x_date(
    limit = c(
      max(Capital_Cities$dt)-(365*5),
      max(Capital_Cities$dt))) +
  ggtitle('Last 05 Years (Capital Cities)')
```



It seems that the temperature completes its cycle every year where it returns to following the pattern, where the first semester is warmer and the second semester starts to cool.

```
## First 10 years (I choose the first 10, because of the data leap
between the 2nd bimester of 1745 to 1750)

ggplot(Capital_Cities,
            aes(x=dt,
                y=AverageTemperature,
                color = City,
                group = City)) +
  geom_line() +
  geom_point() +
  xlab("") +
  theme_ipsum() +
  theme(
    axis.text.x = element_text(angle=60, hjust=1)) +
  scale_x_date(
    limit = c(
      min(Capital_Cities$dt),
      min(Capital_Cities$dt)+(365*10))) +
  ggtitle('First 10 Years (Capital Cities)')
```

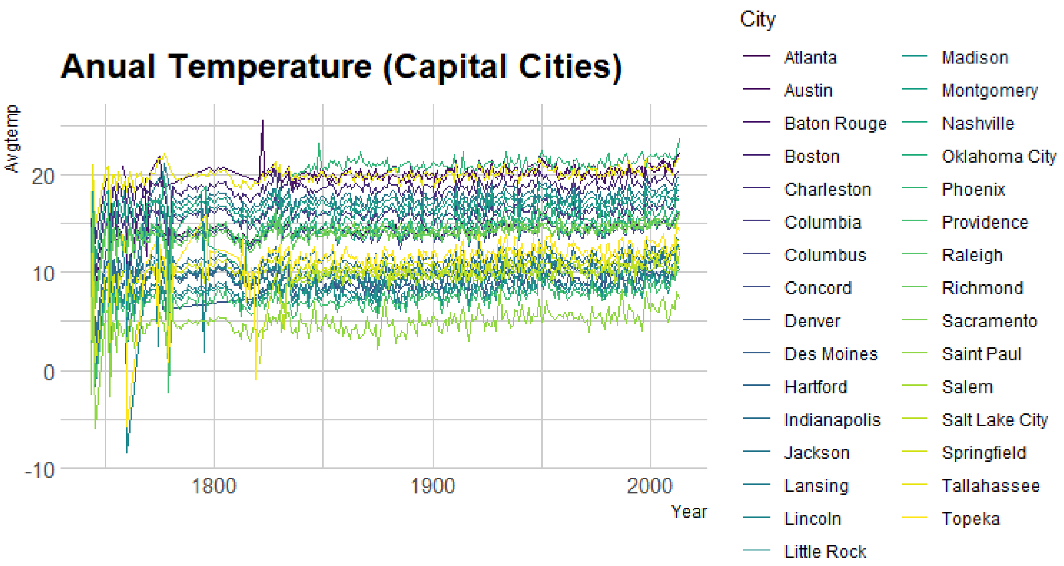


The pattern remains the same, content, we perceive the difference in the amplitude ends between the analyzed periods

In order to check growth over time, let’s do an annual analysis

Stage 7 - Anual Analysis

```
Capital_Cities_Year <- Capital_Cities[, c(2,4,6)]%>%
  group_by(Year, City) %>%
  summarise(Avgtemp = mean(AverageTemperature))

Capital_Cities_Year %>%
  ggplot( aes(x=Year, y=Avgtemp, group=City, color=City)) +
  geom_line() +
  scale_color_viridis(discrete = TRUE) +
  theme(legend.position="none") +
  ggtitle("Anual Temperature (Capital Cities") +
  theme_ipsum()
```



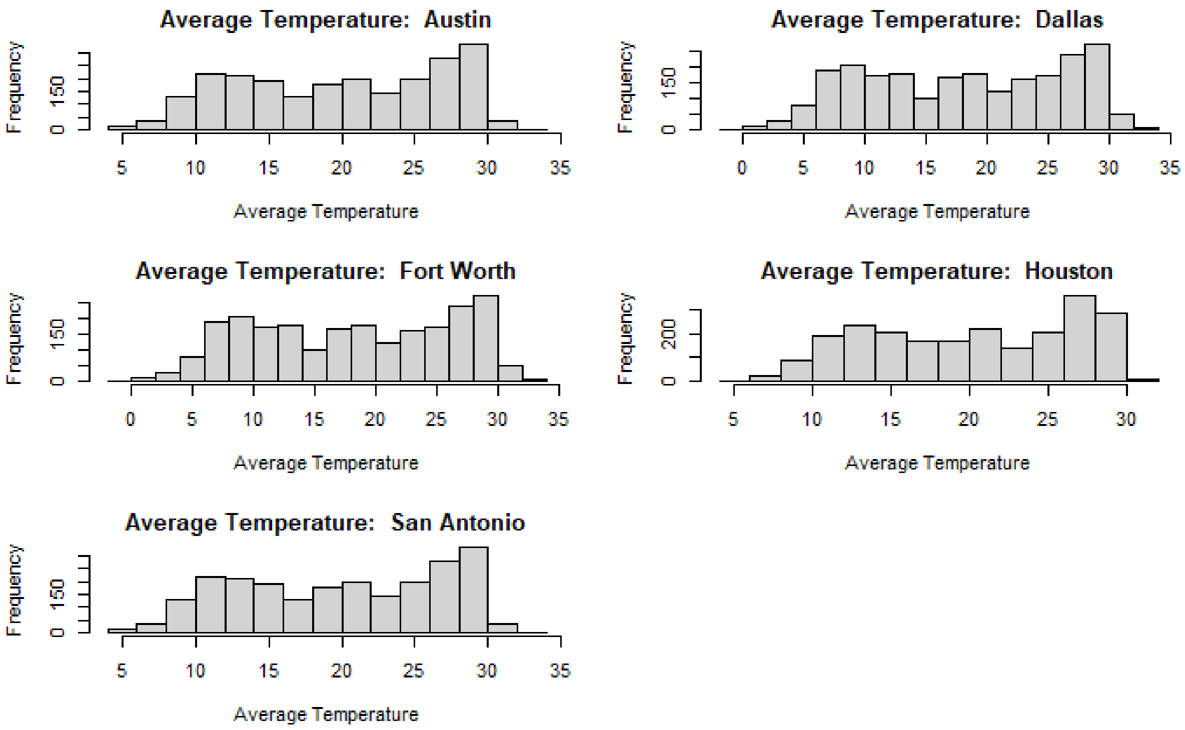
As the years go by, the thermal amplitude decreases, as the temperature gets warmer too

```
range(Capital_Cities_Year$Avgtemp[Capital_Cities_Year$Year < 1760])

## [1] -5.84250 21.05125

range(Capital_Cities_Year$Avgtemp[Capital_Cities_Year$Year > 1990])

## [1]  4.010833 23.566667

mean(Capital_Cities_Year$Avgtemp[Capital_Cities_Year$Year < 1760])

## [1] 10.79774

mean(Capital_Cities_Year$Avgtemp[Capital_Cities_Year$Year > 1990])

## [1] 14.14854
```

Despite the thermal amplitude not having changed so much, it is possible to conclude that in recent times the average temperatures have been rising, with a difference of approximately 3.5º

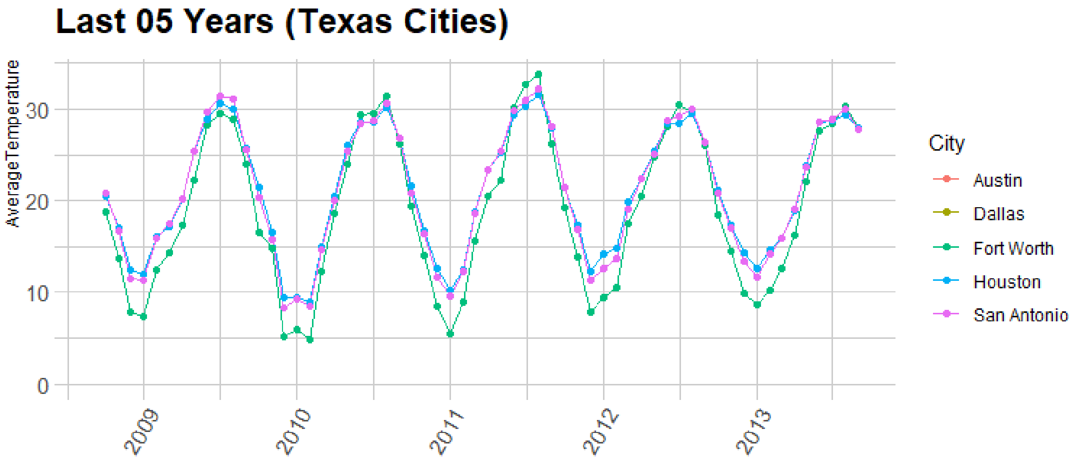
Stage 8 - Texas

```
## Now, let's compare the results with some Texas cities
Texas_Cities <- c('Austin','Dallas', 'Fort Worth', 'Houston', 'San Antonio')

par(mfrow=c(3,2))
for (cities in Texas_Cities) {
  hist(
    US_Climate$AverageTemperature[US_Climate$City == cities],
    main = paste("Average Temperature: ", cities),
    xlab = "Average Temperature"
  )
}
par(mfrow=c(1,1))
```

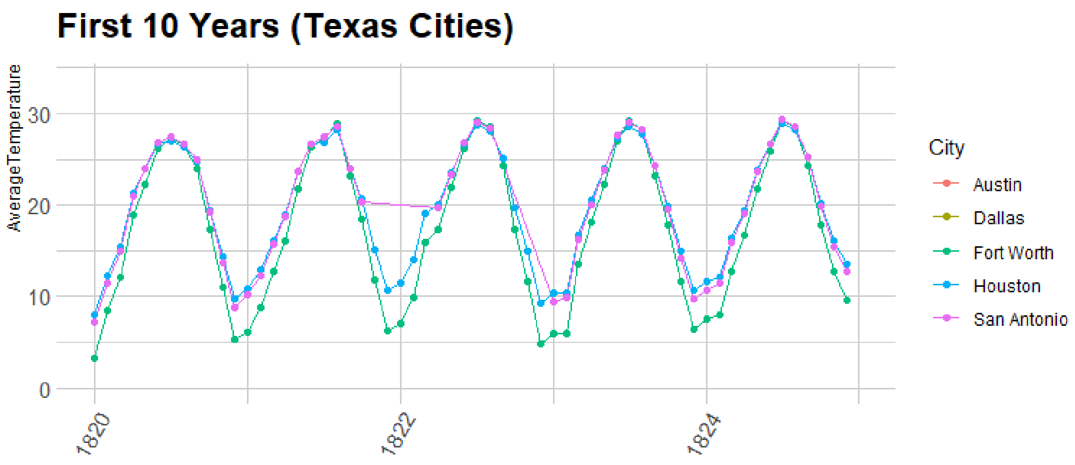


Stage 9 - Analysis over time (Texas)

```
## create a dataframe for texas cities to analyze dates

Texas_Cities_df <-  US_Climate[US_Climate$City %in% Texas_Cities,]
Texas_Cities_df$dt <- as.Date(Texas_Cities_df$dt)

## Last 05 year analysis
ggplot(Texas_Cities_df,
       aes(x=dt,
           y=AverageTemperature,
           color = City,
           group = City)) +
  geom_line() +
  geom_point() +
  xlab("") +
  theme_ipsum() +
  theme(
    axis.text.x = element_text(angle=60, hjust=1)) +
  scale_x_date(
    limit = c(
      max(Texas_Cities_df$dt)-(365*5),
      max(Texas_Cities_df$dt)))+
  ggtitle('Last 05 Years (Texas Cities)')
```



```
## First 05 year analysis
ggplot(Texas_Cities_df,
       aes(x=dt,
           y=AverageTemperature,
           color = City,
           group = City)) +
  geom_line() +
  geom_point() +
  xlab("") +
  theme_ipsum() +
  theme(
    axis.text.x = element_text(angle=60, hjust=1)) +
  scale_x_date(
    limit = c(
      min(Texas_Cities_df$dt),
      min(Texas_Cities_df$dt)+(365*5))) +
  ggtitle('First 10 Years (Texas Cities)')
```



Stage 10 - Analyze Texas Cities

```
## Statistics

for (city in Texas_Cities) {
  print(city)
  print(
    summary(Texas_Cities_df$AverageTemperature[Texas_Cities_df$City ==
city]))
}

## [1] "Austin"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.02   13.78   20.36   20.00   26.70   32.17
## [1] "Dallas"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -0.07   10.76   18.28   18.09   25.87   33.74
## [1] "Fort Worth"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -0.07   10.76   18.28   18.09   25.87   33.74
## [1] "Houston"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.51   14.41   20.60   20.25   26.57   31.52
## [1] "San Antonio"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.02   13.78   20.36   20.00   26.70   32.17
```

As we have seen, the state of Texas continues to increase temperatures over the years, as well as having an average temperature higher than the average of the country by about 5º to 7º

Conclusion

As you can see, the general temperature of the United States has been between its 15 - 25 degrees celsius. The cities of Austin, Dallas and Houston follow a similarity in temperature, as they are close together, while New York is more sparse, ranging from 25 to -5 degrees.

Disclaimer:

```
## Disclaimer: a good part of this project was largely done in the Data
Science Academy, Big Data Analytics with R and Microsoft Azure Machine
Learning course (part of the Data Scientist training)
```

End