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 - Collecting the Data

Here is the data collection, in this case a csv file downloaded from http://berkeleyearth.org/data

```r
## Packages
library(readr)
library(dplyr)
```

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

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

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

```r
library(ggplot2)
library(scales)
```

```
##
## Attaching package: 'scales'

## The following object is masked from 'package:readr':
##
##     col_factor
```

```r
library(data.table)
```

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

```
library(tibble)

# Collecting the Data
df <- fread("GlobalClimate.csv")
```

Stage 2 - Exploring the Data

```
## 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.

US_Climate <- subset(
  df, Country == "United States",
  select =
    c(dt,
      AverageTemperature,
      AverageTemperatureUncertainty,
      City))
```

Stage 3 - Processing and cleaning data

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

```
## [1] 0.07497574
```

```
## We have 7% 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
```

```
str(US_Climate)
```

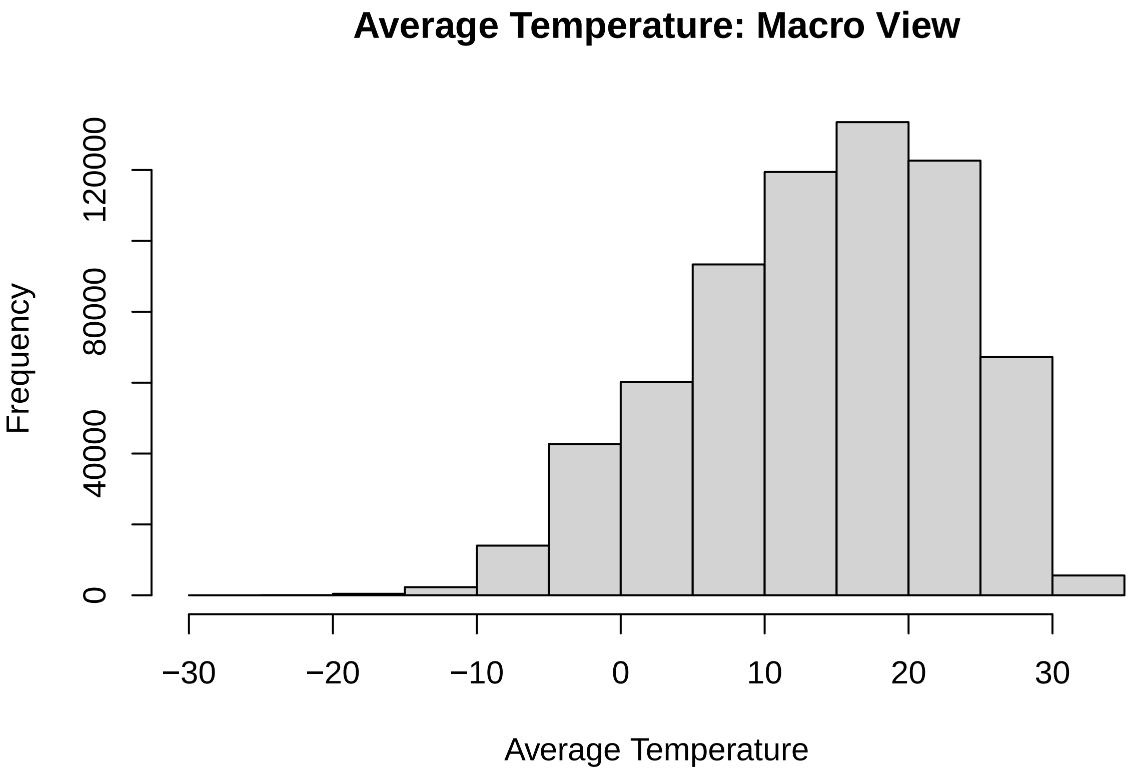
```
## Classes ’data.table’ and ’data.frame’:   661524 obs. of  4 variables:
##  $ dt                           : IDate, format: "1820-01-01" "1820-02-01" ...
##  $ AverageTemperature           : num  2.1 6.93 10.77 17.99 21.81 ...
##  $ AverageTemperatureUncertainty: num  3.22 2.85 2.4 2.2 2.04 ...
##  $ City                         : chr  "Abilene" "Abilene" "Abilene" "Abilene" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

Stage 4 - Preparating and Organizating Data

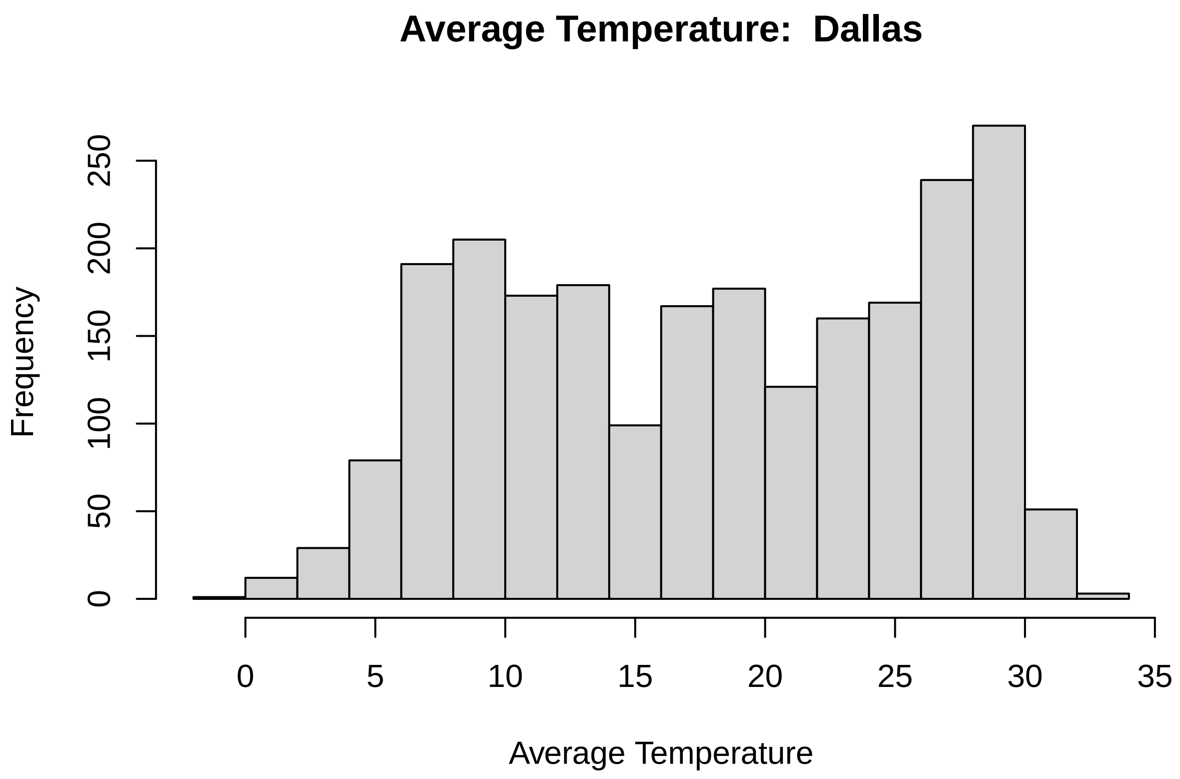
```r
## Converting Data and creating Month and Year column
US_Climate$dt <- as.POSIXct(US_Climate$dt, format = "%Y-%m-%d")
US_Climate$Month <- month(US_Climate$dt)
US_Climate$Year <- year(US_Climate$dt)
```

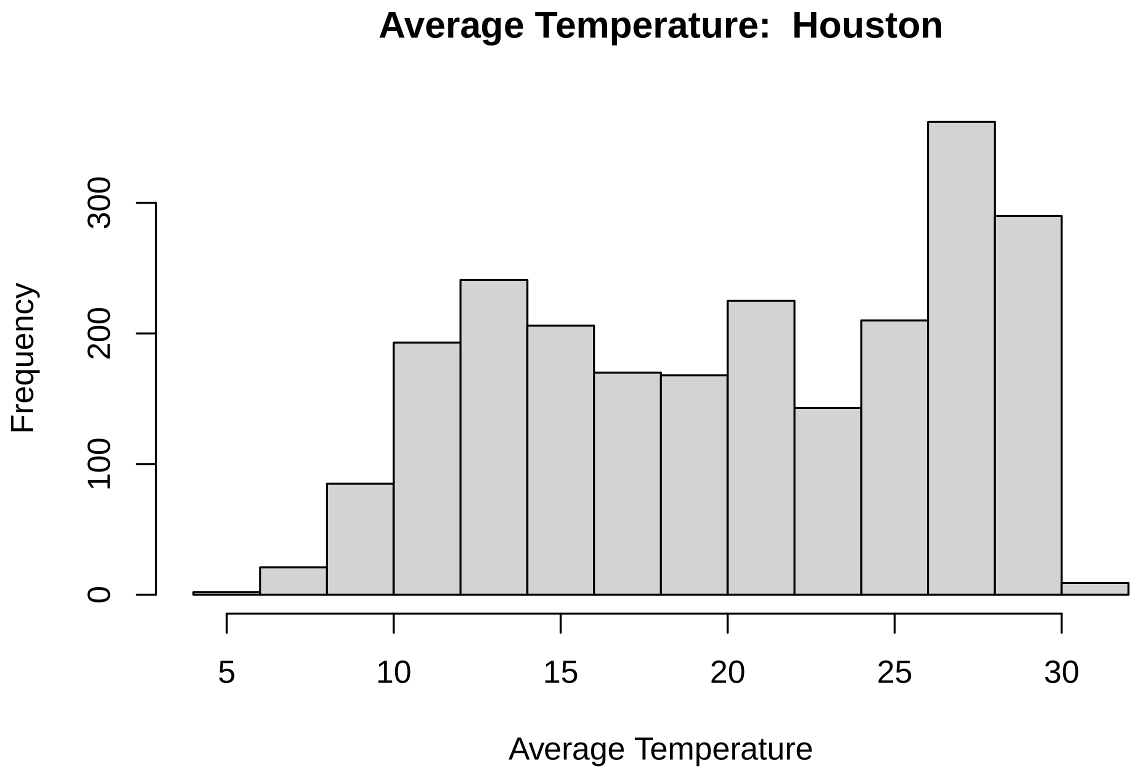
Stage 5 - Macro View

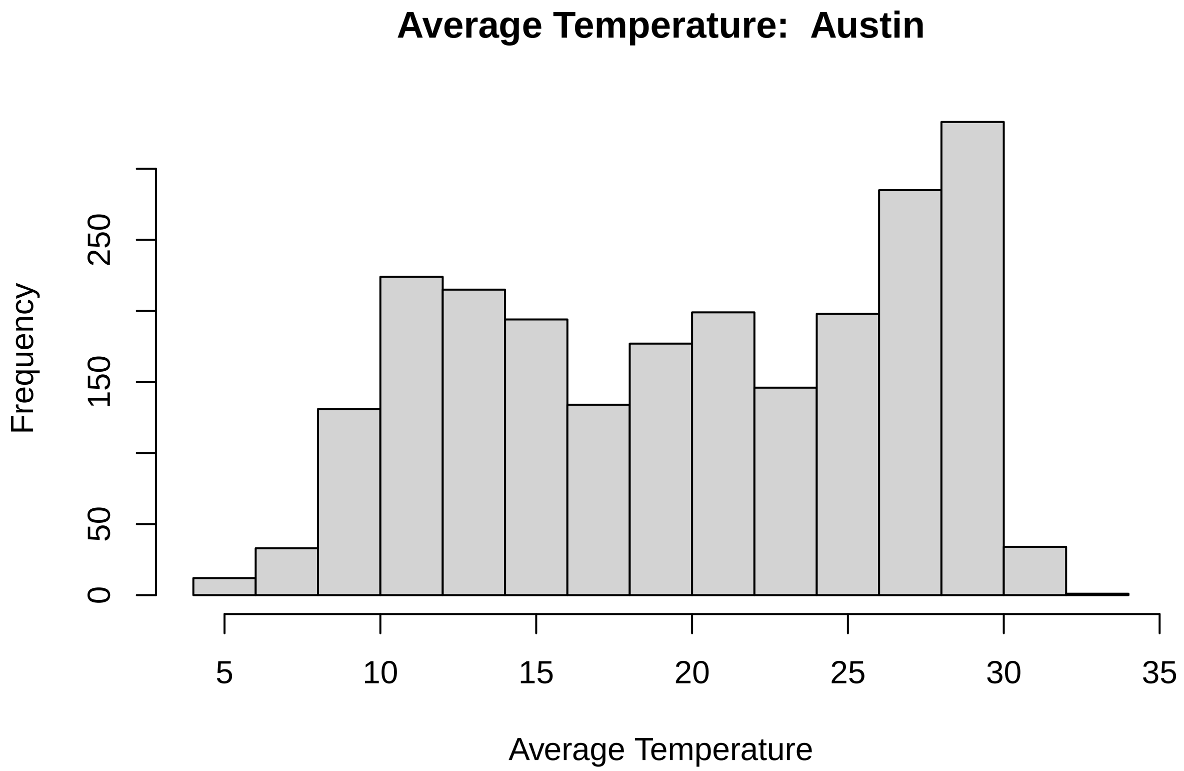
```r
## Macro View
hist(
  US_Climate$AverageTemperature,
  main = "Average Temperature: Macro View",
  xlab = "Average Temperature"
  )
```

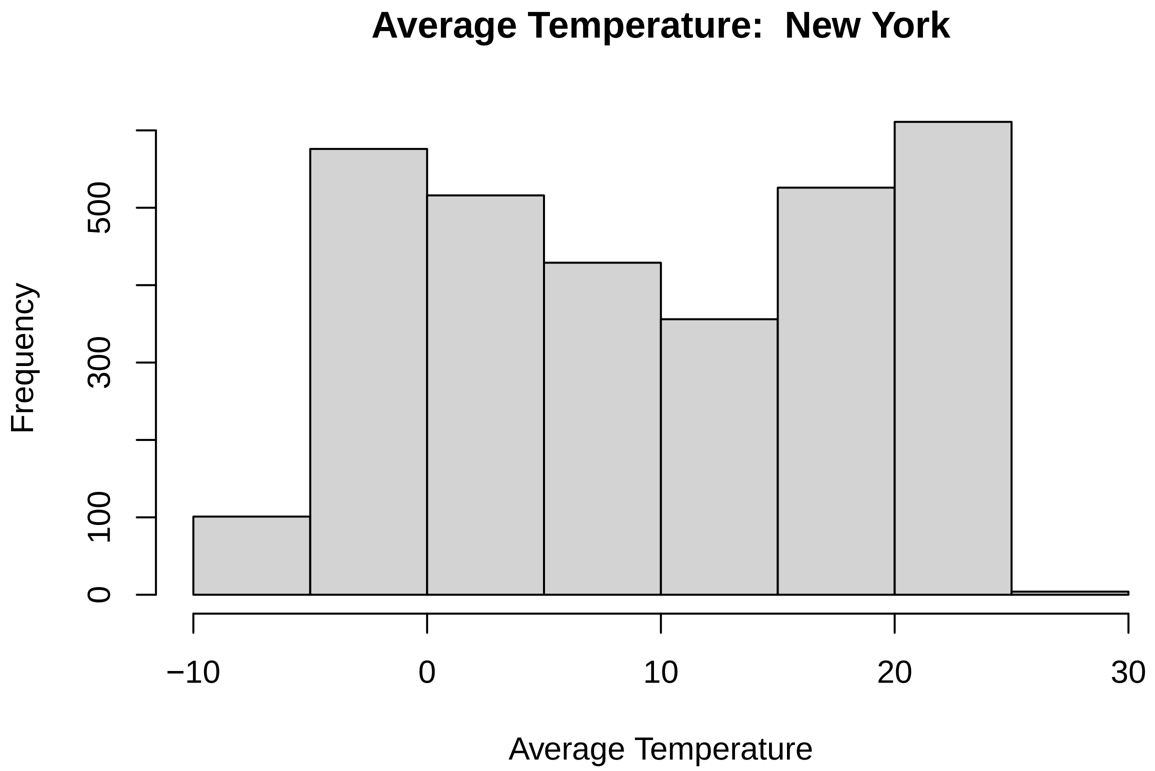


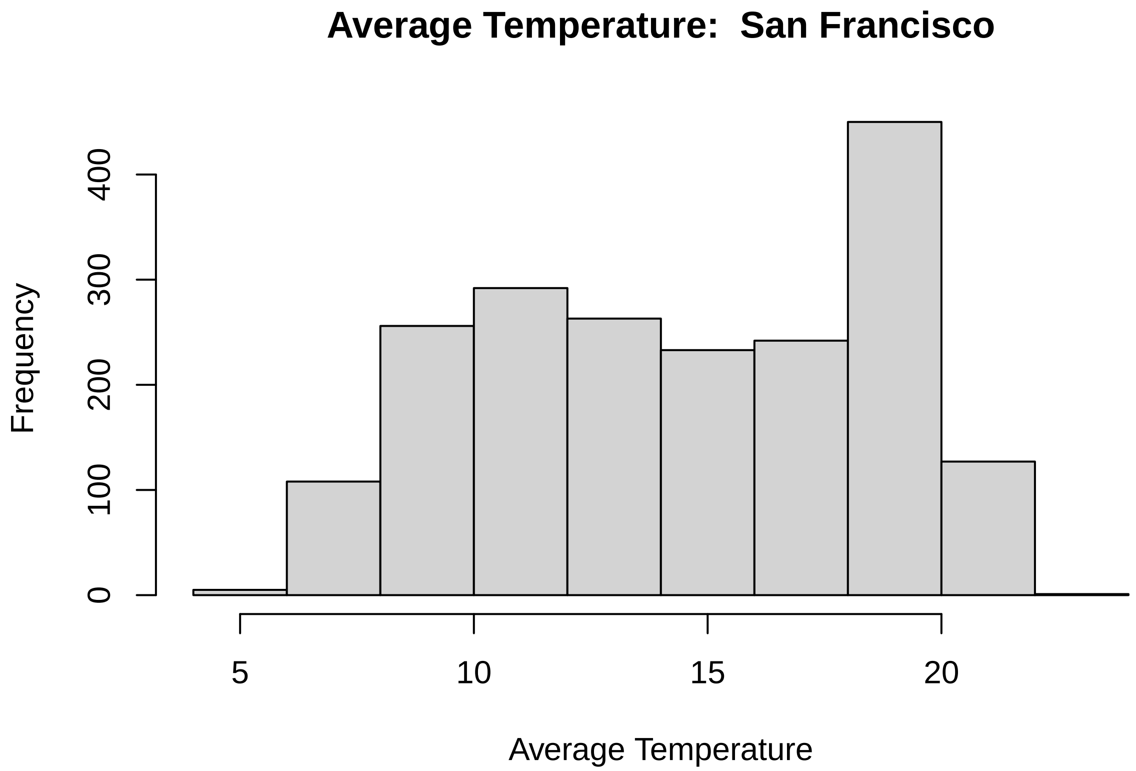
```r
## Let's see some cities
x <- c('Dallas','Houston', 'Austin', 'New York', 'San Francisco', 'Chicago')

for (n in x) {
  hist(
    US_Climate$AverageTemperature[US_Climate$City == n],
    main = paste("Average Temperature: ", n),
    xlab = "Average Temperature"
  )
}
```

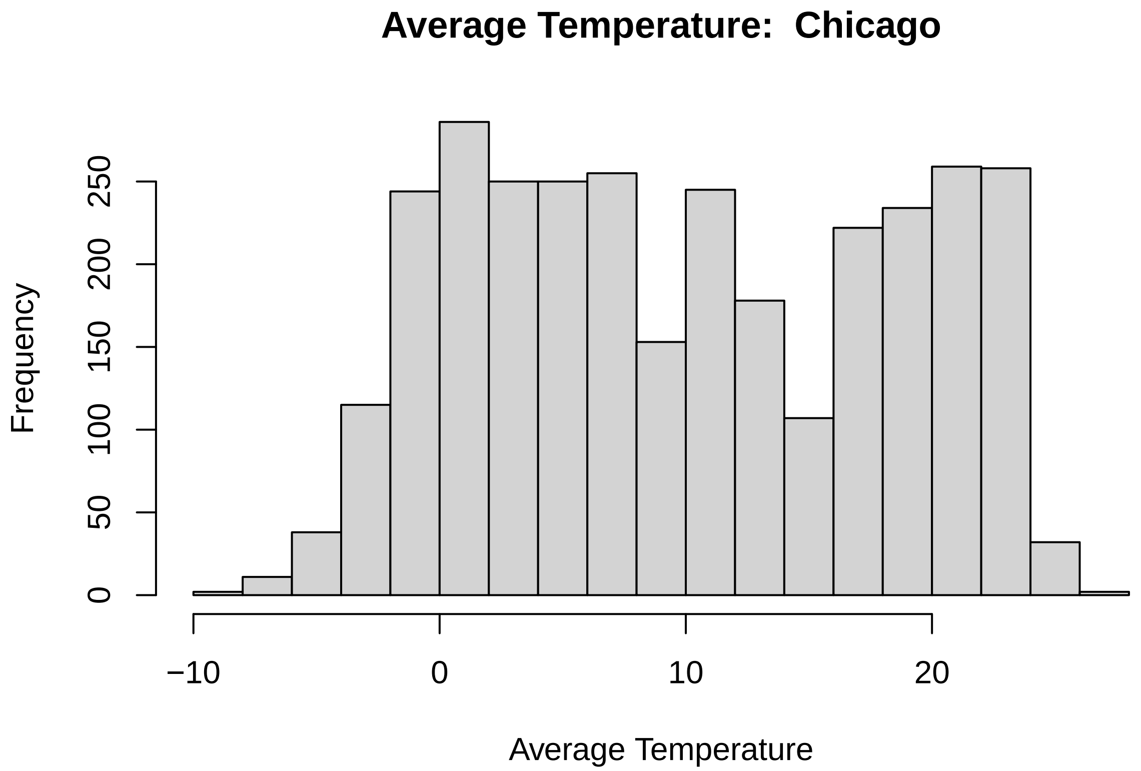



Average Temperature: Houston
Frequency
0
100
200
300
5
10
15
20
25
30
Average Temperature


Average Temperature: Austin
Frequency
0
50
150
250
5
10
15
20
25
30
35
Average Temperature


Average Temperature: New York
Frequency
0
100
300
500
−10
0
10
20
30
Average Temperature


Average Temperature: San Francisco
Frequency
0
100
200
300
400
5
10
15
20
Average Temperature



Conclusion

As you can see, the general temperature of the United States has been between its 15 - 25 degrees celsius. The cities of Austin, Texas 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)
```

Fim

www.datascienceacademy.com.br